



**This electronic thesis or dissertation has been  
downloaded from Explore Bristol Research,  
<http://research-information.bristol.ac.uk>**

*Author:*

**Tsai, Kuei-Ju**

*Title:*

**A multidimensional investigation of a data-driven approach to learning collocations**

**General rights**

Access to the thesis is subject to the Creative Commons Attribution - NonCommercial-No Derivatives 4.0 International Public License. A copy of this may be found at <https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode>. This license sets out your rights and the restrictions that apply to your access to the thesis so it is important you read this before proceeding.

**Take down policy**

Some pages of this thesis may have been removed for copyright restrictions prior to having it been deposited in Explore Bristol Research. However, if you have discovered material within the thesis that you consider to be unlawful e.g. breaches of copyright (either yours or that of a third party) or any other law, including but not limited to those relating to patent, trademark, confidentiality, data protection, obscenity, defamation, libel, then please contact [collections-metadata@bristol.ac.uk](mailto:collections-metadata@bristol.ac.uk) and include the following information in your message:

- Your contact details
- Bibliographic details for the item, including a URL
- An outline nature of the complaint

Your claim will be investigated and, where appropriate, the item in question will be removed from public view as soon as possible.

# **A Multidimensional Investigation of a Data-Driven Approach to Learning Collocations**

**Kuei-Ju Tsai**

**A dissertation submitted to the University of Bristol in accordance with the  
requirements of the degree of Doctor of Philosophy in the Faculty of  
Social Sciences and Law**

**Graduate School of Education**

**2011**

## **Abstract**

This study investigates a corpus-assisted pedagogical approach to developing the collocational knowledge of English as a Foreign Language (EFL) learners - a data-driven approach to learning collocations (DALC). To gain a full understanding of such an approach, three key dimensions thereof are explored: the learning product, learning processes and learner perceptions of DALC.

186 undergraduate EFL learners in Taiwan participated in the study. The participants came from four intact classes, two of which were randomly assigned to an experimental group, and the other two were assigned to a control group: the former received DALC intervention, while the latter encountered target collocations through teacher instruction. Measurements of learners' collocational knowledge were taken from both groups before and after DALC intervention (or non-intervention). Collocational knowledge was examined at three levels: receptive and controlled productive knowledge (as measured by collocation tests), and free productive knowledge (as measured by the collocations used in writing assignments). To understand how collocation learning occurred with DALC, the thinking processes in which learners engaged as they undertook the DALC task were elicited with a mentalistic measure (concurrent think-aloud) and a behaviouristic measure (parallel corpus queries). A questionnaire was administered to elicit learners' perceptions of DALC.

The findings indicate that DALC had a positive impact on learners' receptive, controlled productive and free productive collocational knowledge. In addition, the participants seemed to have an increased awareness of the usage-based and language-specific nature of collocations. The quantitative and qualitative changes in the learners' collocational knowledge may be attributable to the intense cognitive processing in which they engaged during the DALC task, as evidenced by a rich array of cognitive and metacognitive strategies employed to approach the task. Generally, the participants held a positive attitude toward DALC, but they were nonetheless concerned about the efficiency of such an endeavour.

Learners' performance, process and perception data provide evidence that DALC is a promising pedagogical approach in developing EFL learners' collocational knowledge and raising collocational awareness. This warrants further research to explore the possibilities and develop the potential of corpus resources in assisting the learning of collocations.

## **Acknowledgments**

I am deeply indebted to my supervisors, Dr. Helen Woodfield, Professor Richard Kiely and Dr. Guoxing Yu. This thesis would not have been possible without their valuable advice, insightful comments and kind support.

I owe a debt of gratitude to the students and teachers who participated in this study.

I am also grateful to my colleagues at the Graduate School of Education and fellow members of the Centre for Research on Language and Education (CREOLE) for assisting me with data coding and offering helpful comments.

I would like to acknowledge the financial support of the Ministry of Education in Taiwan.

My sincere thanks go to Chong Tao and Pei-Ying Chen for their friendship and encouragement throughout.

Finally, and most importantly, I would like to express my profound gratitude to my parents, my brother and Shun-Chien, who inspired me deeply and supported me wholeheartedly, and whom I love very much.



## Author's Declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: 蔡奎如 Kuei-Ju Tsai DATE: 04/05/11

## Table of Contents

Abstract.....	i
Acknowledgments.....	ii
Author's Declaration .....	iii
Table of Contents .....	iv
List of Abbreviations.....	viii
List of Figures .....	ix
List of Tables.....	x
List of Appendices.....	xii
 CHAPTER 1 INTRODUCTION.....	 1
1.1 Contextualizing collocation in the EFL classroom .....	1
1.2 The pedagogical use of corpora .....	3
1.3 Research aims .....	5
1.4 Structure of the thesis .....	7
1.5 Summary.....	9
 CHAPTER 2 LITERATURE REVIEW .....	 10
2.1 From a rule-based view of language to a usage-based view of language .....	10
2.2 Defining collocations .....	12
2.2.1 Contextualizing collocations.....	12
2.2.2 Characterizing collocations.....	14
2.2.2.1 Habitual co-occurrence of words.....	15
2.2.2.2 Arbitrariness in the choice of collocates .....	16
2.2.3 Classifying collocations.....	18
2.2.3.1 Classification criteria .....	18
2.2.3.2 Classification taxonomy.....	22
2.3 A corpus-assisted approach to developing collocational knowledge – DALC .....	25
2.3.1 The rationale for developing collocational knowledge .....	26
2.3.2 The rationale for DALC.....	28
2.3.2.1 Authenticity versus genuineness .....	31
2.3.2.2 Profusion .....	33
2.3.2.3 Learner autonomy.....	36
2.4 How collocation learning occurs with DALC – Theoretical underpinnings from cognitive and sociocultural perspectives.....	38
2.4.1 Cognitive approaches to SLA .....	38
2.4.1.1 Learning as establishing connections - Connectionism .....	39
2.4.1.2 Learning as processing information – Information processing approaches .....	42
2.4.2 Sociocultural theory.....	46
2.4.2.1 Mediation and appropriation .....	46
2.4.2.2 Tools for mediation.....	48
2.4.3 Combining the best of two worlds – Accounting for DALC with cognitive and sociocultural perspectives .....	49
2.5 Empirical evidence of corpus-assisted language learning.....	51
2.5.1 The product of corpus-assisted vocabulary learning.....	52
2.5.1.1 Effects on definitional knowledge .....	52

2.5.1.2	Effects on collocational knowledge .....	54
2.5.2	The processes of corpus-assisted language learning .....	56
2.6	Summary .....	60
<b>CHAPTER 3 METHODOLOGY .....</b>		<b>62</b>
3.1	Philosophical stance .....	62
3.1.1	Ontological and epistemological commitment to constructivism .....	63
3.1.2	Methodology informed by positivism .....	65
3.2	Research aims and design .....	67
3.3	Participants .....	69
3.4	DALC intervention .....	70
3.5	Collecting and analyzing data on learning product .....	73
3.5.1	Collecting data on collocational knowledge .....	73
3.5.2	Instruments eliciting data on collocational knowledge .....	74
3.5.2.1	Collocation tests .....	74
3.5.2.2	Writing assignments .....	78
3.5.2.3	Collocation learning worksheets .....	79
3.5.3	Analyzing data on collocational knowledge .....	80
3.5.3.1	Marking tests and worksheets .....	80
3.5.3.2	Analyzing tests and worksheets .....	81
3.5.3.3	Identifying collocations in writing assignments .....	82
3.5.3.4	Coding and analyzing collocations in writing assignments ....	83
3.6	Collecting and analyzing data on thinking processes .....	87
3.6.1	Collecting data on thinking processes during corpus consultation .	87
3.6.1.1	Rationale for think-aloud .....	88
3.6.1.2	Validity issues of think-aloud .....	89
3.6.1.3	Piloting think-aloud .....	91
3.6.1.4	Conducting think-aloud in the main study .....	92
3.6.2	The worksheet for think-aloud .....	93
3.6.3	Analyzing think-aloud protocols and parallel corpus queries .....	93
3.7	Collecting and analyzing data on learner perceptions .....	94
3.7.1	Collecting data on learner perceptions .....	95
3.7.2	The questionnaire eliciting learner perceptions .....	95
3.7.2.1	Developing the questionnaire .....	95
3.7.2.2	Piloting the questionnaire .....	97
3.7.3	Analyzing questionnaires .....	98
3.8	Validity and reliability of the research .....	99
3.9	Ethical issues .....	100
3.10	Summary .....	102
<b>CHAPTER 4 EVALUATING LEARNING PRODUCT OF DALC .....</b>		<b>104</b>
4.1	Effects of DALC on the receptive and controlled productive knowledge of collocations .....	104
4.1.1	Entry-level knowledge .....	105
4.1.2	Learner performance on collocation tests .....	107
4.1.3	Changes in the receptive and controlled productive knowledge of collocations .....	109

4.2	Effects of DALC on the free productive knowledge of collocations.	113
4.2.1	The number of collocations.....	114
4.2.2	The types of collocations .....	116
4.2.3	The acceptability of collocations.....	120
4.2.4	Miscollocations .....	123
4.2.4.1	Interlingual factors.....	126
4.2.4.2	Intralingual factors.....	130
4.3	Formative evaluation.....	133
4.4	Summary.....	135
CHAPTER 5 EXPLORING LEARNING PROCESSES OF DALC.....		137
5.1	Corpus consultation behaviour .....	137
5.1.1	Corpus query methods .....	138
5.1.2	Additional reference tools.....	140
5.2	Cognitive strategies.....	142
5.2.1	Deduction.....	144
5.2.2	Induction .....	146
5.2.2.1	Grouping and differentiating .....	148
5.2.2.2	Inferencing.....	149
5.2.3	Substitution .....	161
5.2.3.1	De-lexicalized words .....	161
5.2.3.2	Synonyms .....	162
5.2.4	Translation .....	163
5.3	Metacognitive strategies.....	165
5.3.1	Planning .....	166
5.3.2	Monitoring .....	168
5.3.2.1	Monitoring at the linguistic level.....	168
5.3.2.2	Monitoring at the level of task management.....	170
5.3.3	Evaluation .....	172
5.4	Summary.....	173
CHAPTER 6 EXAMINING LEARNER PERCEPTIONS OF DALC.....		176
6.1	Demographic information.....	176
6.2	English learning resources.....	177
6.3	Vocabulary learning experience.....	178
6.3.1	Lexico-syntactic relations of words.....	179
6.3.2	Perceived importance of components of word knowledge .....	180
6.3.3	Deductive and inductive approaches to learning vocabulary.....	182
6.4	Awareness of collocations .....	183
6.5	Perceptions of DALC .....	184
6.5.1	Hands-on experiences of DALC .....	184
6.5.1.1	Perceived level of difficulty of corpus data.....	184
6.5.1.2	The amount of corpus data .....	186
6.5.1.3	Perceived authenticity .....	188
6.5.1.4	The overall perception of DALC.....	190
6.5.2	Perceived benefits of pedagogical use of corpora.....	191
6.5.2.1	Vocabulary learning .....	191
6.5.2.2	Language learning.....	194
6.5.3	Perceptions of web-based concordancers .....	195
6.5.3.1	Concordancer features.....	196



	6.5.3.2	User preferences for concordancers.....	197
6.6		Summary.....	199
CHAPTER 7 CONCLUSION.....			201
7.1		Key findings and discussion .....	201
	7.1.1	Learning product of DALC.....	201
	7.1.2	Learning processes of DALC.....	204
	7.1.3	Learner perceptions of DALC.....	206
7.2		Limitations and directions for future research .....	209
	7.2.1	Measuring the free production of collocations.....	209
	7.2.2	Additional factors influencing corpus-assisted language learning.. .....	210
	7.2.3	Strengthening the product-process link with case studies .....	211
	7.2.4	Triangulating questionnaire with follow-up interview .....	212
	7.2.5	Diachronic investigation into corpus-assisted collocation learning .....	213
7.3		Pedagogical implications.....	213
	7.3.1	Developing collocational knowledge and raising collocational awareness .....	214
	7.3.2	Scaffolding corpus-assisted language learning .....	215
	7.3.2.1	Developing corpus consultation skills .....	215
	7.3.2.2	Raising critical awareness of the strengths and limitations of corpus resources.....	216
	7.3.2.3	Fostering autonomous learning .....	217
7.4		Concluding remarks .....	218
References.....			220
Appendices.....			234

## **List of Abbreviations**

<b>BAAL</b>	<b>British Association for Applied Linguistics</b>
<b>BNC</b>	<b>British National Corpus</b>
<b>CALL</b>	<b>computer-assisted language learning</b>
<b>CMC</b>	<b>computer-mediated communication</b>
<b>DALC</b>	<b>data-driven approach to learning collocations</b>
<b>DDL</b>	<b>data-driven learning</b>
<b>EFL</b>	<b>English as a Foreign Language</b>
<b>FC</b>	<b>free combination</b>
<b>FE</b>	<b>fixed expression</b>
<b>FL</b>	<b>foreign language</b>
<b>L1A</b>	<b>first language acquisition</b>
<b>L2</b>	<b>second language</b>
<b>LLS</b>	<b>language learning strategy</b>
<b>MI</b>	<b>mutual information</b>
<b>MWU</b>	<b>multi-word unit</b>
<b>NNS</b>	<b>non-native speaker</b>
<b>NS</b>	<b>native speaker</b>
<b>POS</b>	<b>part-of-speech</b>
<b>RC</b>	<b>restricted collocation</b>
<b>SCT</b>	<b>socio-cultural theory</b>
<b>SD</b>	<b>standard deviation</b>
<b>SLA</b>	<b>second language acquisition</b>
<b>TEL</b>	<b>technology-enhanced learning</b>
<b>TL</b>	<b>target language</b>
<b>UG</b>	<b>Universal Grammar</b>
<b>VN</b>	<b>verb-noun</b>

## List of Figures

Figure 2-1	Phraseological categories.....	23
Figure 2-2	Composite/formal units along continua.....	23
Figure 2-3	Concordances of 'weight' from the Collins WordbanksOnline English Corpus .....	29
Figure 2-4	Concordances of 'tantamount' from the Corpus of Contemporary American English.....	35
Figure 3-1	The data collection procedure.....	68
Figure 3-2	Concordances of 'highly' from the Collins WordbanksOnline English Corpus .....	72
Figure 4-1	The miscollocation profile of the experimental group .....	125
Figure 5-1	Query methods and corresponding search scope .....	139
Figure 5-2	A screenshot of 'Yahoo Knowledge' .....	141
Figure 6-1	The perceived hierarchy of components of word knowledge.....	181
Figure 6-2	Perceived usefulness of corpus consultation for components of word knowledge.....	193
Figure 6-3	Degrees of preference for the designated concordancers .....	198



## List of Tables

Table 2-1	Word knowledge framework in Nation (2001).....	13
Table 2-2	The parallel between sets of collocation classification criteria .....	20
Table 2-3	Types of lexical collocations in Benson, Benson and Ilson (1997) .....	25
Table 2-4	Definitions of language learning strategies .....	43
Table 2-5	Taxonomy of LLSs in O'Malley and Chamot (1990).....	44
Table 3-1	The titles and genres of the writing assignments.....	78
Table 3-2	The criteria for marking gap-filling questions.....	80
Table 3-3	The working classification of lexical collocations .....	83
Table 3-4	The coding scheme for miscollocations .....	87
Table 3-5	A summary of research methods.....	102
Table 4-1	Descriptive statistics for the pretest .....	105
Table 4-2	Between-group comparison of pretest scores.....	105
Table 4-3	Descriptive statistics for the pretest and post-test.....	107
Table 4-4	Within-group comparison of test scores (experimental group) .....	108
Table 4-5	Within-group comparison of test scores (control group) .....	108
Table 4-6	Changes in the scores for MC questions and gap-filling questions .....	110
Table 4-7	Collocation production in writing assignments .....	114
Table 4-8	Within-group comparisons of the number of lexical collocations per 100 words .....	115
Table 4-9	Descriptive statistics for the types of collocation used in writing assignments.....	116
Table 4-10	Collocation production profiles across studies.....	118
Table 4-11	Within-group comparisons of the proportions of FCs and RCs in writing assignments.....	120
Table 4-12	Within-group comparison of the proportions of RCs in control group's writing assignments.....	120
Table 4-13	Descriptive statistics for the miscollocations in writing assignments ..	121
Table 4-14	Within-group comparisons of collocational acceptability rate.....	121
Table 4-15	Descriptive statistics for the miscollocations in the writing assignments of the experimental group.....	123
Table 4-16	Descriptive statistics for collocation learning worksheets .....	134
Table 5-1	Definitions of the cognitive strategies identified in this study .....	144
Table 5-2	A segment of S5's verbal protocols and parallel queries .....	144
Table 5-3	A segment of S15's verbal protocols and parallel queries .....	147
Table 5-4	A segment of S4's verbal protocols and parallel queries .....	150
Table 5-5	A segment of S1's verbal protocols and parallel corpus queries.....	155
Table 5-6	S1's queries when approaching Question 1-1 .....	171
Table 6-1	Learner experiences of consulting reference tools for collocational information .....	178
Table 6-2	Learner awareness of lexico-syntactic relations of words .....	179
Table 6-3	Perceived importance of components of word knowledge .....	180
Table 6-4	Preferred approaches to learning vocabulary .....	182
Table 6-5	Awareness of collocations .....	183
Table 6-6	Perceived level of difficulty of corpus data.....	185
Table 6-7	Perceptions of the amount of corpus data .....	187
Table 6-8	Perceptions of corpus language .....	188
Table 6-9	Overall perceptions of DALC.....	190

Table 6-10 Perceived benefits of corpus consultation on vocabulary learning.....192

Table 6-11 Perceived benefits of corpus consultation on language learning .....195

Table 6-12 Perceived accessibility of web-based concordancers.....196

Table 6-13 Perceived importance of concordancer features .....197

## List of Appendices

Appendix 1: Consent form for the experimental group .....	234
Appendix 2: Consent form for the control group.....	235
Appendix 3: Consent form for the think-aloud participants.....	236
Appendix 4: The handout for DALC training session .....	237
Appendix 5: The designated web-based concordancers.....	239
Appendix 6: The pilot test .....	241
Appendix 7: The pretest for the main study .....	245
Appendix 8: The post-test for the main study.....	249
Appendix 9: The worksheets for DALC intervention sessions .....	253
Appendix 10: The worksheet for the think-aloud session .....	258
Appendix 11: The pilot questionnaire .....	260
Appendix 12: The questionnaire for the main study (Chinese version).....	263
Appendix 13: The questionnaire for the main study (English version) .....	266
Appendix 14: Descriptive statistics for questionnaire responses.....	269
Appendix 15: Restricted collocations in writing assignments (experimental group) .....	272
Appendix 16: Restricted collocations in writing assignments (control group) .....	273
Appendix 17: Miscollocations in writing assignments (experimental group).....	275
Appendix 18: Miscollocations in writing assignments (control group).....	277
Appendix 19: Inferential statistics for Questionnaire Items 20, 21 and 50 .....	279

## Chapter 1 Introduction

### 1.1 Contextualizing collocation in the EFL classroom

Over the past twenty years, the field of second language acquisition (henceforth SLA) has seen a gradual departure from a *generative, rule-based* view of language learning, towards an *emergentist, usage-based* one. Central to usage-based models is actual language use, or *exemplars*, ‘that are present in the input that forms the basis of complex patterns and from which regularities emerge’ (Gass and Selinker, 2008:220), as opposed to abstract language rules that characterize rule-based accounts. Language, in the *emergentist, usage-based* view, is composed of a considerable proportion of *formulaic language*, which accounts for one-third to one-half of any discourse, as evidenced in empirical work such as Erman and Warren (2000) and Foster (2001), albeit the proportion varies with genre, register or mode, amongst other variables. As Weinert (1995:184) observes, ‘linguistic studies have begun to show that ready-made language sequences which cannot easily be accounted for in terms of an economical system of generative rules may be more pervasive than generally acknowledged’. Given its ubiquity in language, *formulaic language* plays a significant role in language comprehension and production. The notion of *formulaic language* is widely recognized, but the terminology thereof much less so. Wray (2000) identified 48 different terms in the literature denoting this linguistic phenomenon, of which *collocation* is one.

In its simplest form, *collocation* is ‘the company words keep together’ (Firth, 1957:195). It follows that collocation is the frequent lexical association between words, and thus, part of what constitutes lexical knowledge, among other components such as meaning, form and pronunciation (Nation, 2001). In principle, collocation is an integral part of what constitutes vocabulary knowledge; in practice, vocabulary tends to be taught as discrete units with only marginal attention paid to collocational associations in some English as a foreign language (henceforth EFL) classrooms (Bahns and Eldaw, 1993; Howarth, 1998a). As an EFL learner myself, many vocabulary lessons in my learning context, Taiwan, have only gone so far as to teach *word meaning*, *word form (spoken and written)* and *grammatical function*: rarely has *lexical collocation* been given due attention. As a consequence of learning words in



isolation, I sometimes feel the need to construct an utterance from scratch every time I wish to communicate, which renders such language production inefficient and error-prone. In such EFL contexts, to compound the problem, not only formal tuition but also learning environment has little to offer in developing collocational knowledge because of the paucity of naturalistic exposure to the target language (henceforth TL) necessary for forming and strengthening collocational links. Martin (1984) notes that *collocational errors*, one of the major dissonances between a lexical item and its appropriate use<sup>1</sup>, are common for second or foreign language (L2) learners, because these learners do not have the exposure needed to generalize typical collocates of a lexical item. Given the lack of naturalistic exposure to the TL, it seems unrealistic to teach vocabulary in isolation in the classroom and to expect L2 learners to notice and pick up collocations through subsequent incidental encounters.

A body of empirical work has demonstrated that many EFL learners have been plagued by limited collocational knowledge, even those at advanced levels. Bahns and Eldaw (1993) found that German EFL learners' collocational knowledge lagged far behind their overall vocabulary knowledge; Zhang (1993) found that non-native writers fell short of the *collocational accuracy* and *variety* attained by their native counterparts; based on the International Corpus of Learner English (ICLE), Granger (1998) compared the use of prefabricated patterns in academic essays written by advanced EFL learners and those by native speakers of English. She found that learners used significantly fewer prefabricated patterns in *number* and in *variety*, compared with their native-speaker counterparts. Also, learners' *production* of prefabricated patterns relied heavily on first language (L1) transfer; as regards *recognition*, learners had an underdeveloped and even partly misguided sense of the *salience* of prefabricated patterns.

While the empirical literature has found evidence of L2 learners' underdeveloped collocational knowledge, the theoretical literature has documented considerable discussion on the positive effects of developing collocational knowledge in language learning. It is generally accepted that collocation is a key aspect of lexical knowledge (Henriksen, 1999; Laufer, 1997; McCarthy, 1990; Nation, 2001; Richards, 1976;

---

<sup>1</sup> According to Martin (1984), four dissonances between a word and its context are *stylistic*, *syntactic*, *collocational* and *semantic dissonances*.

Schmitt and Meara, 1997; Wesche and Paribakht, 1996). A developed sense of collocational knowledge is essential to transfer receptive lexical knowledge into productive use. Moreover, having a fair amount of formulaic sequences (e.g., collocations) at one's disposal saves cognitive processing time and effort, thus enhancing *fluency* in language production (Nattinger and DeCarrico, 1992; Pawley and Syder, 1983). Theoretical hardliners, most notably connectionists (e.g., N. Ellis, 2003), go on to claim that collocational knowledge is language knowledge *per se*, because symbolic representations (e.g., grammar patterns) are derived from a substantial number of exemplar formulaic sequences. Given the above evidence of EFL learners' underdeveloped collocational knowledge and the scholarly discussion on the importance of collocational knowledge in language learning and use, the theoretical and empirical literature both call for pedagogical actions to raise learners' awareness of collocations and develop their collocational knowledge.

## 1.2 The pedagogical use of corpora

The advent of computer technology has revolutionized the theory and practice of language education worldwide. Its potential to provide easy access to a wealth of information makes it a powerful tool for language study as well as language learning. Computer-assisted language learning (henceforth CALL) is now more than just a buzzword, it has increasingly become part of the pedagogical landscape in the language classroom. Writing about computer applications in the field of SLA, Chappelle (2001:1) notes:

[a]s we enter the 21<sup>st</sup> century, everyday language use is so tied to technology that learning language through technology has become a fact of life with important implications for all applied linguists, particularly for those concerned with facets of second language acquisition (SLA).

Various types of computer technology have been applied to language teaching and learning, including multimedia learning software, corpora and concordancers, mobile learning, computer-mediated communication (CMC, e.g., web conferencing, blogging), to name but a few. Notably, there has been a growing interest in the use of corpus resources to assist L2 teaching and learning, because such resources provide users with a plethora of language in use which many L2 classrooms fail to offer (Bernardini, 2004). A corpus, according to Sinclair (1996 cited in Granger, 2002), is 'a collection of pieces of language that are selected and ordered according to explicit

linguistic criteria in order to be used as a sample of the language'. Modern computerized corpora are compiled from tens or hundreds of millions of running words in a broad range of written and/or spoken texts. These mega-size corpora along with user-friendly concordancing software make huge collections of real language data available at one's fingertips. Recognizing the profound impact of corpora on language study and learning, Hunston (2002:1) states, '[i]t is no exaggeration to say that corpora, and the study of corpora, have revolutionized the study of language, and of the applications of language, over the last few decades'. Corpora have long been used by linguists for language description and analysis, and the findings have accordingly informed the development of teaching materials (e.g., textbooks, dictionaries, grammar books). It is not until recently that corpus as a pedagogical resource has begun to receive increasing attention within the field of SLA, as Fox (1998:43) puts it, 'the use of concordances in the classroom is in its infancy as a language teaching technique'.

Johns (1991, 1994) has brought the pedagogical use of corpora into prominence by coining the term *data-driven learning* (henceforth DDL) to describe a pedagogical approach characterized by learners' direct exposure to, and analysis of, language data in corpora. His work has proved extremely influential in advocating learners' self-directed observation and analysis of real language data in corpora. Johns (1991) argues that learners need to be guided to observe and analyze the TL in corpora, in much the same way as corpus linguists discover facts about that language. The considerable appeal of corpora to language professionals lies in their ability to 'make the invisible visible' (Tribble and Jones, 1990:11) by aligning multiple contexts of occurrence of a particular form. In other words, the profusion of aligned concordances makes the collocates frequently preceding or following a target word salient. Such saliency of lexical patterning is otherwise difficult, if not impossible, to emerge through incidental encounters over time in contexts where naturalistic exposure to the TL is lacking. This feature of corpora fits well with the formulaic nature of collocations, thus making a corpus a valuable resource for observing collocational patterns and raising awareness of collocability. In view of the volume of real language data in corpora, corpus resources can be usefully exploited to aid collocation learning in FL contexts where naturalistic exposure to the TL is scarce.



Corpus consultation has been attested to facilitate various aspects of language learning, including vocabulary, grammar, discourse analysis, translation and writing. However, despite a corpus' major strength in highlighting collocational patterns, there are few empirical studies that investigated corpus-assisted collocation learning, with the exception of Sun and Wang (2003) and Chan and Liou (2005). Though these two studies examined the *product* of corpus consultation on the learning of collocations (i.e., grammatical collocations and verb-noun collocations respectively), they did not look into the *process* leading to the claimed learning outcomes, nor the *perceptions* learners had of such a pedagogical approach to learning collocations. While the impact of corpus consultation on collocation learning has proved to be positive, the picture of corpus-assisted collocation learning remains somewhat blurred, because a critical question remains unanswered as to *how* learning takes place through corpus mediation. Furthermore, as an integral part of lexical knowledge, collocational knowledge is not mentally represented in an all-or-nothing, dichotomous manner, but falls on a continuum from the *recognition* end through *controlled production* to the *free production* end. Previous studies have only gone so far as to investigate one or two of the above level(s) of collocational knowledge, so the full extent to which corpus consultation facilitates this aspect of lexical knowledge remains unclear. With all these unanswered questions in mind, I set out to explore collocation learning mediated by corpus consultation.

### 1.3 Research aims

As an EFL learner and teacher, I am acutely aware of the challenges facing other fellow learners – limited awareness and knowledge of collocations – attributed to scarce naturalistic exposure to the TL compounded by teaching and/or learning vocabulary in isolation. Motivated by my personal language learning and teaching experience, I embarked on a journey of researching collocation learning assisted by corpus consultation with the aim of contributing to a deeper understanding of this promising pedagogical approach. Corpus-assisted collocation learning is referred to as the *data-driven approach to learning collocations* (henceforth DALC) in this study (see 2.3.2 for the rationale for replacing Johns' DDL with DALC). The current research was conducted in my learning and teaching context, Taiwan, with the participation of 186 undergraduate EFL learners. To bridge the gap in the current

understanding of DALC (as noted in the previous section), the present study aims to provide a more comprehensive picture thereof, from three key dimensions: *learning product, learning processes* and *learner perceptions*.

The *learning product* of DALC is concerned with the changes (quantitative and qualitative) in the learner's collocational knowledge brought about by such pedagogical mediation. More specifically, the changes that took place in the learner's *receptive, controlled productive* and *free productive knowledge of collocations* were examined to understand the extent to which and the ways in which DALC impacted collocational knowledge. This issue leads to the development of the first research question:

RQ1: Does a data-driven approach to learning collocations facilitate EFL learners' development of collocational knowledge? If so, how does it facilitate such development?

Flowerdew (1996:112) draws attention to 'a paucity of critical perspectives in concordancing literature', arguing that much has been written on *what* can be done with corpora in language learning, but relatively little on *how* learners actually fare with such technology-enhanced learning activities. To respond to this call for more in-depth empirical work on *how* learners approach corpus-driven language tasks, the present study examines the *thinking processes* taking place as the learner observes and analyzes corpus data to induce collocational patterns. Although a corpus can make collocational patterns salient through aligned concordances, it does not follow that the human mind will mirror exactly what it is exposed to, in the same way that concordancers do. In other words, the language input provided by corpora is subject to cognitive processing in the mind before it is committed to memory, hence *learning*. It is hypothesized that an underlying mechanism transforms concordances on a computer screen to learners' collocational knowledge. With this assumption in mind, this study aims to illuminate the underlying mechanism accounting for collocation learning mediated by corpus consultation. The second research question asks:

RQ2. What is the nature of the thinking processes EFL learners engage in during the data-driven approach to learning collocations?

In addition, the current research is also concerned with how learners perceive such an innovative pedagogical approach as DALC, which differs markedly from traditional

ones in terms of language input (e.g., quantity, source and presentation mode) and the ways in which such language input is delivered (i.e., induced by the learner rather than imparted by the teacher). Although the theoretical literature has well documented what corpus resources have to offer language learning (as will be reviewed in 2.3.2), it is primarily based on the assertions made by researchers or expert corpus users, and relatively little empirical work has taken into account the perspective of novice corpus users. To gain insights into learners' voice, the study raises the third research question:

**RQ3. How do EFL learners perceive the data-driven approach to learning collocations?**

Through a multidimensional and in-depth exploration of how collocation learning is mediated by corpus consultation, this study aims to provide a sound basis for a wider pedagogical application of corpus resources to assist collocation learning in L2 contexts.

## **1.4 Structure of the thesis**

This thesis consists of seven chapters, each of which is briefly introduced as follows.

**Chapter 2** reviews the literature on defining collocation and developing collocational knowledge. It begins with an observation that the growing interest in collocation stems from a gradual change in the view of language and language learning within the field of SLA. It then contextualizes the linguistic phenomenon of collocation in the area of lexicology. Following this, it sets out the characterizations of collocation from the perspectives of theoretical linguistics and corpus linguistics. Many researchers point out that the major challenge facing collocation research is the lack of a coherent theoretical framework (Carter, 1998; Nation, 2001; Schmitt and Carter, 2004), as evidenced by the wide disagreement on the terminology noted in Section 1.1. To delineate the research focus of the present study, this chapter critically examines the criteria and taxonomies for classifying collocations.

I then turn to the second focus of this chapter, namely the development of collocational knowledge. The rationale for developing collocational knowledge is discussed from different theoretical positions. A corpus-assisted pedagogical approach to learning collocations, namely DALC, is proposed, with a consideration of the



features of corpus resources and the affordance for FL learning. The literature on SLA is reviewed with the aim of providing theoretical underpinnings as to how *collocation learning* occurs with DALC. In addition to the theoretical literature, this chapter also reviews empirical work on the *learning product* (i.e., learning outcomes) and *learning processes* of corpus-assisted language learning.

**Chapter 3** provides a detailed account of the methodology employed to investigate the key dimensions of DALC: the *learning product*, *learning processes* and *learner perceptions*. The data elicitation methods and instruments include tests, worksheets, writing assignments, think aloud interviews, corpus query records and a questionnaire. As both quantitative and qualitative measures have been used, the methods of data analysis include descriptive and inferential statistical analyses, and thematic analysis of think-aloud verbal protocols. The strengths and limitations of such a mixed-method research are critically discussed with reference to validity, reliability and ethical issues.

**Chapters 4-6** detail the findings and discussion in relation to the three research questions. **Chapter 4** addresses the issue of the *learning product* of DALC (RQ1), outlining the *quantitative and qualitative changes* that occurred in learners' collocational knowledge after receiving DALC intervention. It presents evidence of learners' changes in *receptive, controlled productive and free productive collocational knowledge*. While Chapter 4 presents evidence of the learning outcomes of DALC, **Chapter 5** reports on a further probe into the thinking processes underlying learners' corpus exploration behaviour (RQ2). An individual's thinking processes cannot be obtained directly from performance data, so a combination of a mentalistic measure (i.e., think-aloud verbal reporting) and a behaviouristic measure (i.e., corpus query records) is used to reconstruct the thinking processes taking place as learners consult corpus resources. **Chapter 6** discusses the *learner perceptions* of DALC (RQ3) elicited from a questionnaire. This chapter presents findings of questionnaire responses in terms of learners' vocabulary learning experience and resources, awareness of collocations, and their perceptions of corpus as a language learning resource, such as the perceived benefits and the challenges encountered in the course of the DALC task.

**Chapter 7** summarizes the major findings of the study in relation to the research aims.

The strengths and limitations of the study are considered with the aim of offering suggestions and directions for future research. Most importantly, in response to what gives impetus to the research, pedagogical implications are provided.

## **1.5 Summary**

This chapter has briefly defined collocation and outlined the problems of learning collocations in FL contexts. It has also proposed a pedagogical approach assisted by corpus resources (i.e., DALC), and has set out the aims of the research. In the following chapter, I will review the theoretical and empirical literature to lay the groundwork for the current research on the data-driven approach to learning collocations.

## Chapter 2 Literature Review

This chapter begins with an observation that the growing interest in collocation stems from a change in the view of language and language learning within the field of SLA (2.1). It then characterizes collocation from the theoretical and corpus linguistic perspectives, and considers the criteria and taxonomies for classifying collocations (2.2). Shifting the focus from a linguistic perspective to a pedagogical one, Section 2.3 first sets out the rationale for developing L2 learners' collocational knowledge and then discusses a corpus-assisted pedagogical approach to mediating collocation learning, DALC. Section 2.4 goes on to consider the theoretical underpinnings of such an approach within the field of SLA. Section 2.4.1 discusses how cognitive approaches to SLA account for corpus-assisted collocation learning. In Section 2.4.2, I argue that the pedagogical use of corpora is grounded in the notion of *mediation* within sociocultural theory (SCT). Recognizing the epistemological differences between cognitive theory and sociocultural theory of language learning, I conceptualize DALC as an interface that combines the best of the two worlds. Section 2.5 reviews empirical studies on the *learning product* and *learning processes* of corpus-assisted language learning. The theoretical and empirical literature reviewed in this chapter is summarized in Section 2.6.

### 2.1 From a rule-based view of language to a usage-based view of language

In the past decades, applied linguistics has seen a steady move from a *generative, rule-based* view of language toward an *emergentist, usage-based* one. The former, most notably Universal Grammar (henceforth UG), holds that language consists of *principles* and *parameters*: the former is an invariable set of abstract rules which characterize the core grammar of all natural languages, while the latter may vary across languages (Chomsky, 1995). Whereas abstract rules are biologically endowed in the human mind, parameters such as the lexicon need to be acquired. In contrast, the emergentist, usage-based view of language maintains that symbolic representations (e.g., grammar rules) do not exist *a priori* in the mind, but emerge from the generalization of a substantial amount of language *exemplars* (see the



definition in 1.1), namely actual usage in a language. As noted earlier (1.1), language in use is composed of a considerable proportion of *formulaic language*, of which *collocation* is one kind.

In conceptualizing the organization of a language, Sinclair (1991) argues that language as a whole is constructed on two principles: the *open-choice principle* (or the slot-and-filler model) and the *idiom principle*. The former, conceptually akin to the generative, rule-based view of language, sees language texts as comprising of a large number of grammatical structures with slots to be filled in by lexical items, the choice of which is constrained by grammaticality and semantics. Prescriptive grammars are predominantly constructed on the open-choice principle (ibid.). On the other hand, the fact that words do not occur randomly in a text shows that the idiom principle operates alongside the open-choice principle. The idiom principle underlies the fact that language, to a large extent, consists of regular patterns and formulaic stretches of words. The emphasis of the idiom principle on the formulaic nature of language coincides closely with the usage-based view of language. The linguistic phenomenon of collocation is a representation of the idiom principle. Sinclair (ibid.) notes the prevalence of the idiom principle in actual language use, and thus calls for more attention to the formulaic nature of language. In other words, lexically-determined behaviour should be given due attention over that determined by grammars. With the steady shift from a rule-based view of language to a usage-based one, *formulaic language* has begun to receive increasing attention within the field of applied linguistics (see 2.4.1.1 for the implications that the two views of language have on SLA).

From the 1970s, the importance of formulaic language in first language acquisition (L1A) and adult language processing has been brought into sharper focus through the work of linguists such as Wong-Fillmore (1979) and Peters (1983). Since then researchers have begun to take notice of the role of formulaic language in language acquisition and use (e.g., Bolinger, 1976; Coulmas, 1979; Widdowson, 1984; Yorio, 1980). However, the crucial role of formulaic language in SLA has not received due attention until the mid-1980s: it has only begun to gain wider recognition due to the emergence of lexico-grammar inspired by the work of Halliday (1985) and Sinclair (1991), and revealing insights yielded from corpus-based research. Whereas the



notion of formulaic language has been generally accepted, the terminology thereof has been less so: Wray (2000) identified 48 terms used in the literature to denote various forms of formulaic language, such as *collocation*, *fixed expression*, *formulaic sequence*, *lexical(ized) phrase*, *multi-word unit*, *routine formulae*, and so on. The present study uses the term *collocation*, and the following sections are devoted to defining and delineating *collocation* (see 2.2.3.2 for collocation in relation to other types of formulaic language).

## **2.2 Defining collocations**

Recent corpus findings have revealed that a far larger proportion of language use is composed of collocations than was previously imagined. The proportions can be as high as one-third to one-half of any discourse, depending on genre, register or mode (cf. Erman and Warren, 2000; Foster, 2001). In view of this, it is important, albeit difficult, to systematize such a ubiquitous linguistic phenomenon. However, to the best of my knowledge, there is no single and universal definition to encapsulate the immense complexity of collocation. In this section, I seek to develop a working definition of collocation for the present study through a characterization of this linguistic phenomenon.

### **2.2.1 Contextualizing collocations**

In the field of lexicology, it is now widely recognized that knowing a word entails much more than knowing the word meaning and word form. Knowledge of derivation, inflection, grammatical function or collocation all constitute word knowledge. As Schmitt (2008:333) states, '[w]hile it is true that the form-meaning link is the first and most essential lexical aspect which must be acquired, and may be adequate to allow recognition, a learner needs to know much more about lexical items, particularly if they are to be used productively'. Beheydt (1987:57) notes that 'the learner has not really semantized a new word until he knows its morphological, syntactic, and collocational profile as well as its meaning potential'. Many researchers have stressed the multidimensionality of word knowledge (N. Ellis, 1995; Harley, 1995; Henriksen, 1999; Laufer, 1997; Nation, 2001; Richards, 1976; Schmitt and Meara, 1997), among which Richards (1976) is the first attempt to draw attention to such a multidimensional nature of word knowledge. Following Richards (ibid.), Nation

(2001) sets out a comprehensive word knowledge framework (see the table below):

**Table 2-1 Word knowledge framework in Nation (2001)**

<b>Form</b>	Spoken	(R, P)
	Written	(R, P)
	Word parts	(R, P)
<b>Meaning</b>	Form and meaning	(R, P)
	Concept and referents	(R, P)
	Associations	(R, P)
<b>Use</b>	Collocations	(R, P)
	Grammatical functions	(R, P)
	Constraints on use (register, frequency)	(R, P)

Note: R = recognition; P = production.

While the framework distinguishes various aspects of word knowledge, they are by no means distinct, but interrelated with varying strengths of association (Schmitt and Meara, 1997). Furthermore, word knowledge does not exist in an all-or-nothing dichotomous manner, it should be conceptualized as a receptive-productive continuum, along which different aspects of word knowledge fall at various positions. When a learner claims to know a word, s/he may know the form-meaning link of the word but does not necessarily know how to use the word in extended contexts. On the other hand, the learner might have some knowledge about a word by being able to recognize it when given options, even if s/he claims not knowing the word. Following such a conceptualization, Laufer (1998) identifies three levels of word knowledge: passive/receptive, controlled active/productive, and free active/productive knowledge. Vocabulary knowledge should be regarded as *continuous* instead of *dichotomous*. As Færch, Haastrup and Phillipson (1984:100) put it,

[r]ather than make the simplistic opposition between ‘active’ and ‘passive’ vocabulary, we should think of vocabulary knowledge as a continuum between ability to make sense of a word and ability to activate the word automatically for productive purposes.

As illustrated in the word knowledge framework above, collocation is one of the key components of word knowledge. Carter (1988:163) describes collocation as ‘an aspect of lexical cohesion which embraces a “relationship” between lexical items that regularly co-occur’. Collocation, in this sense, is concerned with how a word is typically used in relation to others, it is therefore a kind of *word association* (in its broad sense). In the field of semantics, collocation has been researched as one of the two types of word associations: *syntagmatic* and *paradigmatic associations*. Meara (2009:6) defines *syntagmatic associations* as ‘associations that complete a phrase (syntagm)’, such as:

BRUSH	teeth
BLACK	mark
BANK	robber

On the other hand, *paradigmatic associations* are 'ones in which the stimulus word and the response that it evokes both belong to the same part of speech, nouns evoking nouns, verbs evoking verbs, and so on' (ibid.), for example:

MAN	woman
HOT	cold
TREE	bush

More precisely, *syntagmatic associations* are the collocational relations of a word, while *paradigmatic associations* are lexical relations such as synonymy, antonymy, hyponymy, meronymy, and so forth. In summary, collocation is concerned with the *syntagmatic* associations among words, and is thus an integral part of what constitutes word knowledge. This section has positioned *collocation* within the broad fields of lexicology and semantics. I will narrow the focus to the characteristics and categorizations of collocations in the following sections.

### 2.2.2 Characterizing collocations

Firth (1957) is among the first to conduct systematic studies on collocations, and is therefore regarded as the one who brought the term into prominence in the field of lexicology. Collocation has thus been referred to as a 'Firthian term', as a tribute to Firth's significant contribution to the study of collocations (Nation, 2001). However, Palmer (1933) had used the term *collocation* and written extensively in this area long before Firth. Palmer (ibid.) underscored the importance of treating collocation as a single unit:

each [collocation]... must or should be learnt, or is best or most conveniently learnt as an integral whole or independent entity, rather than by the process of piecing together their component parts.

Nation (2001) summarizes the rationale underpinning Palmer's (ibid.) choice of the term *collocation*:

1. It had been used before in the Oxford English Dictionary in 1750.
2. It had no definite association with other meanings.
3. It was made of Latin parts so it was an international word.
4. It could be used in a variety of disciplinary areas.



In addition, the term *collocation* has also been adopted by a number of authoritative collocation dictionaries such as the *Oxford Collocations Dictionary for Students of English*, *LTP Dictionary of Selected Collocations*, and so on. It is an established term in theory and in practice, so the present study also uses the term *collocation*. Notwithstanding the lack of a universal definition of collocation, there is a general consensus as to the two defining characteristics: the habitual co-occurrence of words and the arbitrariness in the choice of collocates. These characteristics are discussed in the following sections.

### 2.2.2.1 Habitual co-occurrence of words

In its simplest form, collocation is the company words keep together (Firth, 1957). To elaborate on the concept of the *company* of words, Carter (1998:51) defines collocation as:

a term used to describe a group of words which occur repeatedly in a language. These patterns of co-occurrence can be grammatical in that they result primarily from syntactic dependencies or they can be lexical in that, although syntactic relationships are involved, the patterns result from the fact that in a given linguistic environment certain lexical items will co-occur.

For example, the noun *attention* habitually co-occurs with the verb *pay*, hence the collocation *pay attention*. Since the words *pay* and *attention* co-occur so frequently, it follows that the frequency of their co-occurrence will reach a significant level when submitted to statistical computations, indicating that the co-occurrence is not a result of pure chance. Nattinger and DeCarrico (1992:36) define collocations as ‘strings of specific lexical items that co-occur with a mutual expectancy greater by chance’. The mutual expectancy between the constituents of a collocation varies in degree: for instance, the collocation *extenuating circumstances* is highly restricted, such that it does not allow any substitution of its constituents (\**extenuating situations* is unacceptable); other collocations, albeit frequent, are not entirely fixed, for example, *adverse circumstances* can be substituted by *difficult circumstances*.

Related to the feature of habitual co-occurrence is *collocational prosody*: that is, part of a word’s meaning is embedded in its collocations (Firth, 1957). For instance, the word *cause* gains part of its meaning from collocates, such as *cause an accident* or *cause an injury*: the verb *cause* thus gains a negative sense from its noun collocates *accident* or *injury* (Stubbs, 1995b).

To delineate collocations systematically, Carter (1998:58) proposes that '[c]ollocational acceptability can be analyzed using techniques of informant analysis in which the intersubjective intuitions of groups of native language speakers are statistically measured and a line drawn between what can be generally allowed and what cannot'. Carter's (ibid.) idea of operationalizing *collocability* as a statistical tendency is made possible by mega-size computerized corpora and concordance software. A fruitful line of research on collocations is *corpus linguistics*, which investigates linguistic regularities and variations through computerized retrieval and analysis of corpus data (Berry-Rogghe, 1973; De Cock, 1998; Fuentes, 2001; Hunston, 2002; Kennedy, 2003; Moon, 1998; Nesselhauf, 2004; Sinclair, 1991). Researchers following this line generally view collocations in a broad sense: 'any recurrent pairs or groups of words which emerge from the corpus with a greater frequency than could be predicted by their individual frequencies as lexical items' (Bonk, 2001:4). More specifically, Sinclair (1991:170) sees collocation as 'the occurrence of two or more words within a short space of each other in a text'. The strengths of association between the constituents of a collocation are measured through statistical methods such as *mutual information*<sup>2</sup> or *T-score*<sup>3</sup>. Collocation in this view does not differ fundamentally from that of prescriptive linguistics with respect to the habitually co-occurring nature, except the former is based on corpus descriptions rather than linguistic prescriptions. In sum, the habitually co-occurring nature is widely recognized as the dominant characteristic of collocations (Cruse, 1986; N. Ellis, 1997; Howarth, 1998a; McCarthy, 1990; Nation, 2001; Nattinger and DeCarrico, 1992; Pawley and Syder, 1983; Schmitt, 2000; Wray, 2002)

#### 2.2.2.2 Arbitrariness in the choice of collocates

Another distinctive feature of collocations is the *arbitrariness* in the choice of collocates. McIntosh (1961) argues that words have only a certain tolerance of compatibility, or a limited 'range' of collocates. Collocability is arbitrarily subject to actual usage of language, and thus does not always conform to semantic grounds. As Bolinger (1976:6) states, '[a] collocation may involve normal senses of all the words

---

<sup>2</sup> A mutual information score 'is a measure of how strongly two words seem to associate in a corpus, based on the independent relative frequency of the two words' (Hunston, 2002:72).

<sup>3</sup> T-score 'is a measure of certainty of collocation', which also takes corpus size into account (Hunston, 2002:73).

in the string, but without the easy possibility of substituting some other word with the same meaning'. To illustrate, *rancid butter* and *sour cream* are frequent collocations, whereas *\*sour butter* and *\*rancid cream* are grammatically well-formed but unconventional combinations. Although *rancid* and *sour* denote a similar smell or taste, they are typically expected to precede *butter* and *cream* respectively, not interchangeably. The latter combinations are rejected by the arbitrariness of actual usage in the language. The non-substitutability of the constituents in certain collocations, or the mutual expectancy between constituents, stems from the fact that collocability is dictated by actual usage in a language. In van Roey's (1990:46) words, 'the linguistic phenomenon whereby a given vocabulary item prefers the company of another item rather than its "synonyms" because of constraints which are not on the level of syntax or conceptual meaning but on that of usage'. Recent corpus-based language descriptions also attest to the fact that lexical choices are not always accounted for on semantic or syntactic grounds (Smadja, 1989).

Since collocability is primarily usage-based, it follows that it is also language-specific. Many collocations have sociocultural connotations and associations, so they cannot be readily translated from one language to another without losing the intended meanings or causing misunderstanding (Moon, 1997). Even if there are analogous collocations across languages, they are rarely identical in form, meaning or constraints on use. For instance, though the collocation *blow your own trumpet* in English may have conceptual equivalents in Chinese, they take on different forms, such as *blow bulls* (吹牛), and thus cannot be literally translated on a word-for-word basis.

The usage-based and language-specific nature of collocations may cause considerable difficulties for L2 learners, since they do not have as much naturalistic exposure to the TL as is needed to allow them to distinguish typical collocations from unconventional ones. Pawley and Syder (1983:215) identify the problems facing learners with respect to the arbitrariness of collocations as follows:

[i]t is a characteristic error of the language learner to assume that an element in [a linguistic] expression may be varied according to a phrase structure or transformational rule of some generality, when in fact the variation (if any) allowed in nativelike usage is much more restricted.

Echoing Pawley and Syder's (ibid.) concern, Howarth (1998a:37) maintains that the learner might be perplexed 'by the way in which specific collocations might be



predicted by analogy, but are arbitrarily blocked by usage'. The above discussion of collocational restrictions has pedagogical implications for raising the learner's awareness of the usage-based and language-specific nature of collocations, and developing their ability to observe and generalize habitually co-occurring patterns in language.

In sum, lexical behaviour is not as flexible as the generative view of language would argue, but rather formulaic to varying degrees. The formulaic nature of such behaviour is dictated by actual usage in a language, and is thus language-specific. Bearing in mind the above defining characteristics of collocation, I will delineate the collocations under investigation in the following subsections.

### **2.2.3 Classifying collocations**

The probability of certain words co-occurring characterizes collocations, whilst causing problems for rigorous descriptions of this seemingly evident yet elusive linguistic phenomenon. Broadly speaking, the difficulty of researching into formulaic language such as collocation lies in a lack of a coherent theoretical framework (Nation, 2001; Schmitt and Carter, 2004; Weinert, 1995). Definitions, criteria for identification and classification, and even terms adopted all vary with theoretical and methodological stances. More specifically, Nation (2001) considers that a major challenge of collocation research is the lack of a consistent way of classifying collocations. The term *collocation* has, thus far, been used as an umbrella term to refer to a wide range of formulaic sequences. The following subsections consider the sets of criteria proposed to classify collocations and the taxonomies of collocations.

#### **2.2.3.1 Classification criteria**

Read and Nation (2004) summarize five major approaches to identifying and/or classifying formulaic sequences: intuition, corpus analysis, structural analysis, phonological analysis and pragmatic/functional analysis. Native speaker intuition, whilst necessary, has constantly been challenged by corpus findings as being unreliable. Corpus-based approaches identify collocations through statistical measures, including the frequency of occurrence and the strength of association among words (e.g., mutual information or T-score). Pragmatic/functional approaches do not fit well



in this study, because it is concerned with *composite/formal units* instead of *functional expressions* (see 2.2.3.2). Hence, the sets of criteria discussed in this subsection are based primarily on structural analysis, with in mind the emphasis of corpus analysis on the frequency of occurrence and strength of association.

As Fernando and Flavell (1981:19) observe, 'idiomaticity is a phenomenon too complex to be defined in terms of a single property. Idiomaticity is best defined by multiple criteria, each criterion representing a single property'. To identify and classify such a complex phenomenon, multiple criterial features need to be taken into account. Carter (1998:70-72), for example, proposes three criteria to determine the lexicality of fixed expressions: *collocational restriction*, *syntactic structure* and *semantic opacity*. Each criterion lies on a cline from a less fixed end to a more fixed one. On the cline of *collocational restriction* lie unrestricted collocation (e.g., *take a walk*), semi-restricted collocation (e.g., *harbor grudges*), familiar collocation (e.g., *vicious circle*) and restricted collocation (e.g., *pitch black*; and irreversible binomials: *ups and downs*, *hit and miss*). The cline in the *syntactic structure* runs from the flexible end (e.g., *break one's heart*), regular with certain constraints (e.g., *smell a rat*), to the irregular end (e.g., *go one better*, *go it alone*). The cline in the *semantic opacity* runs from transparent (e.g., *long time no see*, *when all is said and done*), semi-idioms/metaphor/idiomatic similes (e.g., *we are all in the same boat*, *as sober as a judge*), semi-transparent (e.g., *the business really take off*) to opaque (e.g., *kick the bucket*). This set of classification criteria is not exclusive to collocations, but applies to all types of lexical patterns. In similar vein, Moon (1997) proposes three criteria for identifying multi-word units (MWUs), including *institutionalization*, *fixedness* and *non-compositionality*. Moon's (ibid.) set of criteria coincides closely with that in Carter (ibid.), notwithstanding the different terms adopted (the parallel is summarized in Table 2-2). *Institutionalization* corresponds to Carter's (ibid.) *collocational restriction* in that it indicates the extent to which a word combination is collocationally restricted or conventionalized in the language. *Fixedness* concerns the extent to which the *syntactic structure* of a word combination is flexible. *Non-compositionality* parallels *semantic opacity*, as both refer to the extent to which the meaning of a word combination can be inferred from the meanings of its constituents. The three criteria above are widely accepted as the basic principles for classifying various types of formulaic sequences.

Nation (2001) expands the three established criteria into ten criterial scales for classifying formulaic sequences (see a visual representation in p. 26). A formulaic sequence may fall into a certain point along each scale based on its degree of the characteristic indicated by that particular scale. Scales 5-9 correspond closely to the sets of criteria established by Carter (ibid.) and Moon (ibid.). The parallel between the three sets is illustrated as follows:

*Table 2-2 The parallel between sets of collocation classification criteria*

Moon (1997)	Carter (1998)	Nation (2001)
Institutionalization	Collocational restriction	Collocational specialization; Lexical fossilization
Fixedness	Syntactic structure	Grammatical uniqueness; Grammatical fossilization
Non-compositionality	Semantic opacity	Semantic opaqueness

In addition to the five overlapping criteria, Nation (ibid.) takes into account five additional criterial features. Scales 2 and 3 are concerned with the relative positions of the components of a word combination in an extended text. Scale 10 centers on the literal meaning or figurative meaning of a word combination. Scale 1 considers the frequency of the components of a word combination co-occurring. Related to the frequency of co-occurrence is the judgment of grammaticality, Scale 4 deals with the degree to which a word combination conforms to grammaticality. For example, the sequence *of the* habitually co-occurs but does not constitute a grammatical chunk. This instance demonstrates that while the frequency of co-occurrence is the primary criterion adopted by corpus linguists to determine if a word combination is a collocation, it is a necessary but by no means sufficient criterion. Although corpus technology is able to extract linguistic regularities from substantial amounts of language data with a speed and rigour that lies beyond the capacities of human brains, identifying and classifying collocations still requires a certain degree of human judgment. For instance, the degree of semantic opacity can only be determined by human judgment. It is nonetheless worth noting that in using the native speaker (NS) intuition measures, it is crucial to ensure intersubjectivity, namely a high degree of inter-rater reliability (Read and Nation, 2004).

## 1. Frequency of co-occurrence

← Frequently occurring together      infrequently occurring together →

## 2. Adjacency

← next to each other      separated by several items →

## 3. Grammatically connected

← grammatically connected      grammatically unconnected →

## 4. Grammatically structured

← well structured      loosely related →

## 5. Grammatical uniqueness

← grammatically unique      grammatically regular →

## 6. Grammatical fossilization

← no grammatical variation      changes in part of speech →

## 7. Collocational specialization

← always mutually co-occurring      all occurring in a range of collocations →

## 8. Lexical fossilization

← unchangeable      allowing substitution in all parts →

## 9. Semantic opaqueness

← semantically opaque      semantically transparent →

## 10. Uniqueness of meaning

← only one meaning      several meanings →

Taken together, a sound approach to identifying and classifying collocations relies as much on formal/structural analysis as on corpus computation and human judgment (this study has used these three measures in identifying and classifying learner collocations, see methodological details in 3.5.3). Hence, a principled measure based on the classification criteria noted above, intersubjectivity check and statistical information (e.g., frequency count, mutual information or t-score) gained from representative corpora is essential to identify and classify collocations.

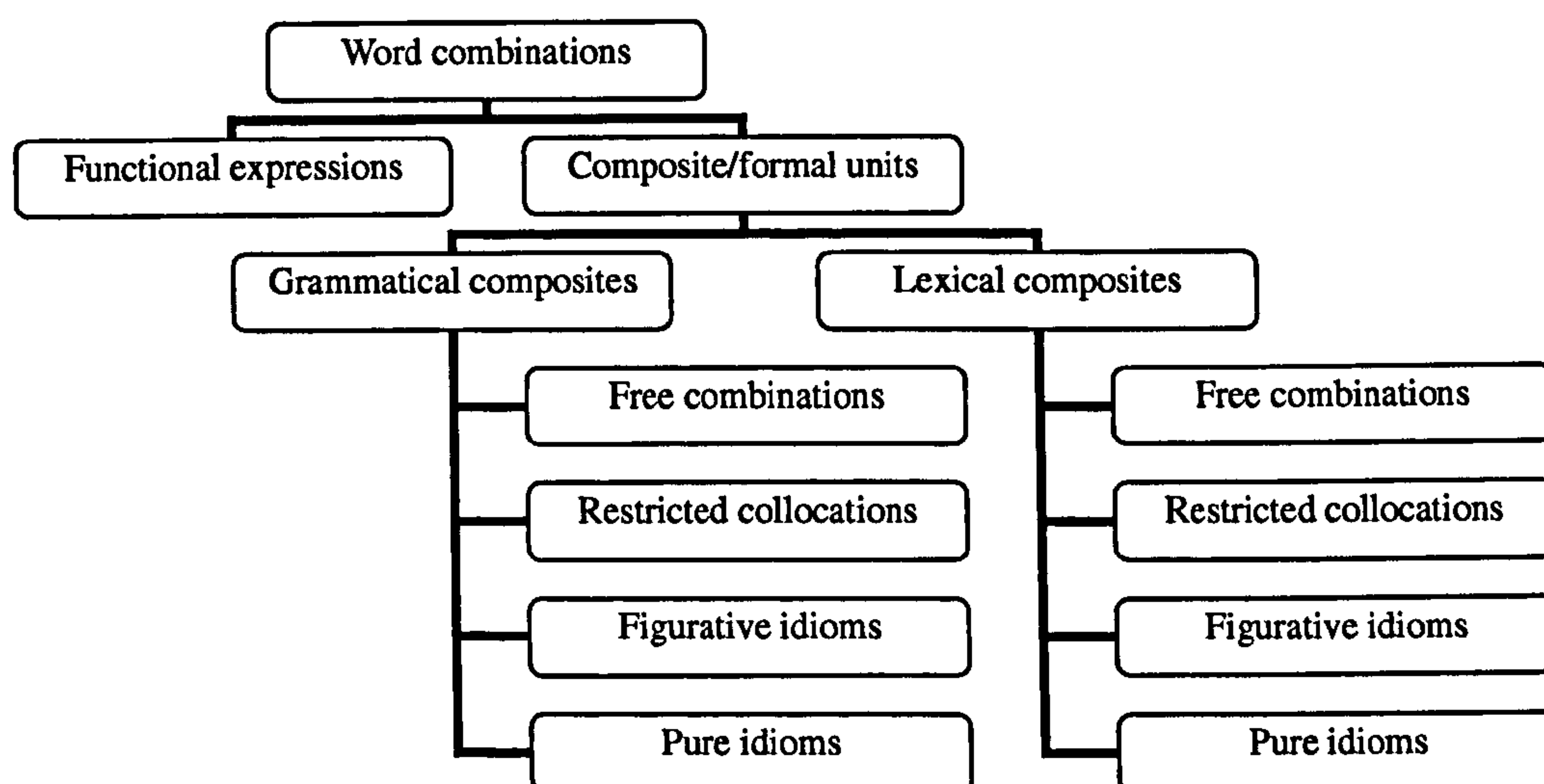
### 2.2.3.2 Classification taxonomy

Wray (2002) adopts the term *formulaic sequences* to encompass a broad range of word combinations, of which *collocation* is one. A *formulaic sequence* is defined as (ibid.:9):

a sequence, continuous or discontinuous, of words or other elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar.

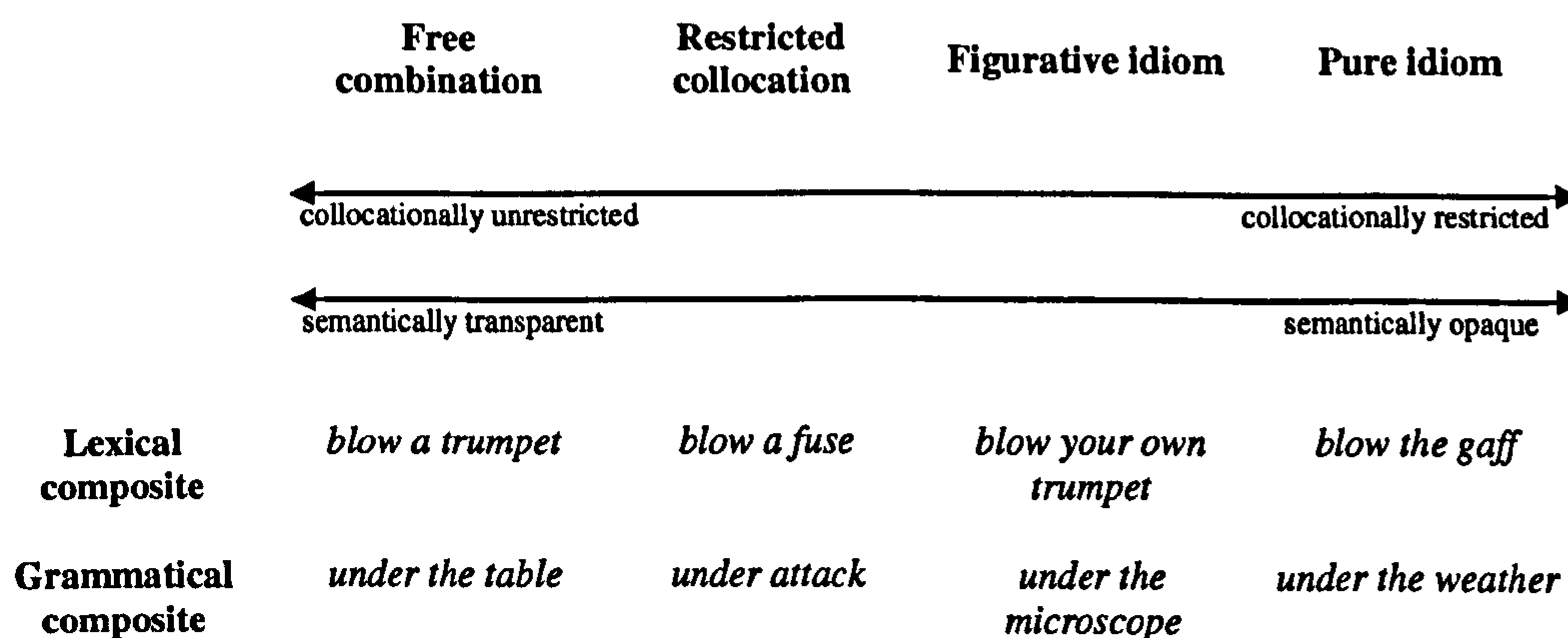
In recent decades, attempts have been made to categorize such a linguistic phenomenon (e.g., Cowie, 1981; Glaser, 1988; Mel'cuk, 1998). In particular, Howarth (1998a) provides the most comprehensive taxonomy which systematically categorizes various forms of formulaic sequences. Drawing on the works in descriptive linguistics (Glaser, 1988), language processing (Bolinger, 1976; Pawley and Syder, 1983) and lexicography (Aisenstadt, 1981; Cowie, 1981), Howarth (ibid.) establishes a taxonomy of word combinations based on criteria such as collocational restrictedness and semantic opacity (see classification criteria in 2.2.3.1). The taxonomy is as follows:





**Figure 2-1** *Phraseological categories*

*Functional expressions* are identified by their functions in discourses, such as *gambits*, *proverbs*, *catchphrases* or *slogans*. *Composite/formal units* have syntactic functions in clauses or sentences, and they are divided into *grammatical composites* and *lexical composites* based on the word class of their constituents. Benson, Benson and Ilson (1986) also distinguish between grammatical and lexical categories, but they focus specifically on collocations rather than word combinations in general (see Benson, Benson and Ilson’s categorization later in this subsection). Based on criteria such as collocational restrictedness and semantic opacity, Howarth (ibid.) distinguishes four subcategories under each composite. The four subcategories are situated on continua from the collocationally unrestricted end to the collocationally restricted one, as well as from the semantically transparent end to the semantically opaque one. See Figure 2-2 below for the subcategories along the continua.



**Figure 2-2** *Composite/formal units along continua*

In terms of collocational restrictedness and semantic opacity, *free combinations* allow free substitution of the constituents, and the constituents reflect their own literal meaning senses. *Restricted collocations* have a limited range of collocates: though they consist of one element used in a restricted, non-literal sense, they are semantically transparent compared with idioms. *Figurative idioms* have a metaphorical meaning as well as a current literal interpretation. *Pure idioms* have a single minimal signification which cannot be separated into semantic constituents, and the meaning cannot be inferred from the meanings of the constituent parts. *Pure idioms* are the most opaque semantically, and are rigid in terms of the substitutability of constituents. In considering the restrictedness of collocation relative to that of free combination and idiom, Bahns (1993:57) characterizes collocation as follows:

The main characteristics of collocations are that their meanings reflect the meaning of their constituent parts (in contrast to idioms) and that they are used frequently, spring to mind readily, and are psychologically salient (in contrast to free combinations).

However, the distinction is not always as clear-cut as expected, as there might be 'transitional areas' (Cruse, 1986) between types of combinations, such as *foot the bill* and *curry favour* which are idiom-like but semantically transparent.

Whereas Howarth (ibid.) subcategorizes grammatical and lexical composites on the basis of collocational restrictedness and semantic opacity, Benson, Benson and Ilson (ibid.) subcategorize grammatical and lexical collocations according to the word class of the constituents. *Grammatical collocations* consist of a content word like verb, noun or adjective along with a grammatical word/structure like preposition, infinitive or clause: for example, *in advance*, *be fond of*. *Lexical collocations* are combinations of content words, such as noun, verb, adjective or adverb: for example, *make a claim*, *ulterior motive*. Benson, Benson and Ilson (1997) identify eight types of grammatical collocations<sup>4</sup> and seven types of lexical collocations (see Table 2-3).

---

<sup>4</sup> Noun + preposition, noun + to + infinitive, noun + that + clause, preposition + noun, adjective + preposition, adjective + to + infinitive, adjective + that + clause, and 19 verb patterns.

**Table 2-3 Types of lexical collocations in Benson, Benson and Ilson (1997)**

Type		Example
verb + noun	CA collocations (consisting of a verb denoting creation and/or activation and a noun/pronoun)	<i>come to an agreement</i>
	EN collocations (consisting of a verb meaning essentially eradication and/or nullification and a noun)	<i>reject an appeal</i>
adjective + noun		<i>weak tea</i>
noun + verb		<i>bees buzz</i>
noun of noun		<i>a pack of dogs</i>
adverb + adjective		<i>deeply moved</i>
verb + adverb		<i>argue heatedly</i>

**Working definition of collocation in this study:** Collocation is the habitual co-occurrence of words, which is grammatically well-formed and semantically transparent. As collocation is largely usage-based and language-specific, the collocability, or mutual expectancy between words, is arbitrary to varying degrees. Drawing on Howarth's (ibid.) taxonomy of phraseological categories (Figure 2-1), this study focuses on *composite/formal units*, particularly *lexical composites*. The learning of *lexical composites* warrants investigations because relatively less pedagogical attention has been devoted to this category compared with *grammatical composites* (Bahns, 1993). More specifically, under the category of *lexical composites*, three types of composites are investigated in this study: *free combinations*, *restricted collocations* and *figurative idioms*. Note that the latter type is referred to as *fixed expressions* in the study to avoid confusion with *pure idioms*. For ease of reporting, the term *collocation* is used interchangeably to denote *free combination*, *restricted collocation* and *fixed expression* unless specified otherwise. In addition to Howarth (ibid.), Benson, Benson and Ilson's (ibid.) classification of *lexical collocations* (see Table 2-3) also serves as the preliminary framework for delineating collocations in this study. The identification and classification of collocations under study were guided by these two frameworks (see 3.5 for methodological details).

## 2.3 A corpus-assisted approach to developing collocational knowledge – DALC

The preceding section has characterized and delineated collocations from a linguistic perspective. This section shifts focus to the *learning* of such a linguistic phenomenon in L2 contexts. As highlighted earlier (1.1), the ubiquity of collocations in language



makes them important for language comprehension and production; nevertheless, the usage-based and language-specific nature renders them difficult for L2 learners. In this section, I will first provide the rationale for developing the collocational knowledge of L2 learners (2.3.1) and then consider a corpus-assisted, pedagogical approach to learning collocations, namely DALC (2.3.2).

### 2.3.1 The rationale for developing collocational knowledge

Scholarly discussion has underscored the importance of developing collocational knowledge in L2 learning from three positions. First, theoretical hardliners such as connectionists hold that collocational knowledge forms the basis of language knowledge, because language is predominantly *usage-based* or *exemplar-based* (N. Ellis, 2003), as opposed to the Chomskyan approaches (Chomsky, 1995) characterized by generative grammar and the creative power of syntactic rules (i.e., *rule-based*). Connectionists/emergentists regard the stored chunks of words as the basis of language knowledge: complex linguistic regularities emerge through repeated exposure to the chunks in language input. The mental representation of language knowledge is therefore *exemplar-based*, as distinct from the *rule-based* representation argued by generative accounts (also see 2.1). This claim is substantiated by a body of empirical evidence which shows that a large proportion of utterances are predictable and even formulaic instead of novel or constructed from scratch every time (N. Ellis, 2003; Erman and Warren, 2000; Foster, 2001). As Bolinger (1976:1) notes,

[o]ur language does not expect us to build everything starting from lumber, nails and blueprint, but provides us with an incredibly large number of prefabs, which have the magical property of persisting even when we knock some of them apart and put them together in unpredictable ways.

This line of argument, albeit hardline, draws attention to the ubiquity of collocations in language, and calls for the need for more pedagogical mediation.

The second strand of thought regards collocational knowledge as an integral part of word knowledge, as noted in Section 2.2.1. Lexical knowledge can be measured in *breadth* and *depth*: the former is vocabulary size, while the latter concerns knowledge of various aspects of a word, such as associations, collocations or constraints on use (see Table 2-1 for components of word knowledge). It is generally accepted that knowing a word entails much more than knowing its form-meaning link (N. Ellis,



1995; Harley, 1995; Henriksen, 1999; Laufer, 1997; Nation, 2001; Richards, 1976; Schmitt and Meara, 1997). Knowing the range of collocates of a word, its contextual preferences and frequency distributions is no less important than other aspects of word knowledge.

The third perspective views collocational knowledge as a major component of language competence, and recognizes the fundamental role of collocations in the comprehension and production of language (Cowie, 1992; Nattinger and DeCarrico, 1992; Pawley and Syder, 1983). For language comprehension, language users rely on collocational knowledge to predict and decode the language input, for example, Underwood, Schmitt and Galpin (2004) found that words are read more quickly when they are part of formulaic sequences than they are in nonformulaic text; Conklin and Schmitt (2008) further confirmed that the processing advantage of formulaic sequences over creatively generated language holds true for both native and non-native speakers. In terms of language production, language users draw on collocational knowledge from their repertoire to construct utterances as appropriate. To be precise, collocational knowledge is important for language production in terms of *accuracy*, *fluency* and *appropriateness* (Pawley and Syder, 1983; Yorio, 1980). *Accuracy* of language production can increase through a store of ready-made chunks at one's disposal (Bonk, 2001). As regards *fluency*, it is widely recognized that a developed sense of collocational knowledge is accompanied by fluency in language production, because ready-made chunks save speakers processing time and effort for encoding and decoding language, and lighten the load of memorization (Moon, 1997; Nation, 2001; Nattinger and DeCarrico, 1992; Pawley and Syder, 1983; Schmitt and Carter, 2004). Without a developed sense of collocational knowledge, learners would have to construct utterances from scratch when communicating (Foster, 2001; Skehan, 1998) or express their meanings with longer but less precise expressions, therefore being more prone to error (Hill, 2000). With more collocations at the learner's disposal, his/her language processing time can be reduced, and fluency enhances accordingly. With respect to *appropriateness*, adequate collocational knowledge helps learners use appropriate collocations as the genre or register demands (Howarth, 1998a; Pawley and Syder, 1983).

The ability to choose natural and idiomatic utterances from a wide range of

grammatically acceptable but highly marked paraphrases is referred to by Pawley and Syder (1983) as *nativelike selection*. Many non-native speakers produce grammatical but unconventional utterances for a lack of a developed sense of collocational knowledge (Chenoweth, 1995). In their seminal work Pawley and Syder (1983) argue that *nativelike fluency* and *nativelike selection* are the two ultimate goals of learning collocations. However, it is problematic to assume native speaker usage as the language norm, not only because the so-called 'native speakers' of English are comprised of speakers from various speech communities and cultural backgrounds, but also because their uses of English vary depending on a wide range of factors such as the context of use, the purpose of communication, the speaker's educational background or socio-economic status, and so on (cf. Davies, 2003). As Carter (1998:39) points out, '[w]hat is "normal" and possible will always be a matter of stylistic choice and relative to a dynamic and negotiable interactive context'. Furthermore, EFL learners do not learn English like L1 learners in terms of the learning environment and resources available to them (see 2.4.3 for discussion on the differences in L1 and L2 learning), native speaker competence does not seem to be a reasonable learning goal for SL/FL learners at large (Seidlhofer, 2005). Though holding reservations about aiming for *nativelike fluency* and *nativelike selection* as the ultimate goal of learning collocations, I nonetheless recognize the importance of collocational knowledge in extending vocabulary knowledge, facilitating language comprehension and enhancing language production. In view of the importance of developing collocational knowledge, I proposed a pedagogical approach to learning collocations, namely DALC.

### 2.3.2 The rationale for DALC

Since collocations are recurring patterns that are ubiquitous in a language, it is sensible to make use of repositories of language in use to facilitate collocation learning, hence the pedagogical application of corpora. While corpora have been providing a sound basis for language studies for decades, their potential for assisting language learning has only begun to be seen since the 1990s. Johns (1991, 1994) has brought the pedagogical use of corpora into prominence by coining the term 'data-driven learning' (DDL) to describe a pedagogical approach which is characterized by the learner's direct access to, and analysis of, the language data in

corpora. By querying and retrieving sentences in which the query word is embedded, the learner can observe how the word is typically used in the language, its frequency of occurrences and contextual preferences. The act of consulting corpora is called *concordancing* (Flowerdew, 1996:97), which is defined as:

a means of accessing a corpus of text to show how any given word or phrase in the text is used in the immediate contexts in which it appears. By grouping the uses of a particular word or phrase on the computer screen or in printed form, the concordancer shows the patterns in which the given word or phrase is typically used.

Figure 2-3 is a sampled concordance of the word *weight*. The alignment of the query word in the middle allows the user to observe the frequent collocates preceding or following the word, in this case, *lose* or *put on*.

very little. Women who exercised put on less	weight	than those who did not. Women under 35, who
his physical activity dwindled and he put on	weight.	It was not an environment conducive to the
pill that will make it easy for us to lose	weight.	[p] Just a day or so ago, some of their
pregnancy and although she put on a lot of	weight,	she was confident that Tim still adored
the national average in maintaining their new	weight.	It also helps if you are a nature lover.
Many women, consciously or not, put on	weight	to avoid marital sex. Weight gain usually
wouldn't we would be putting a lot of	weight	on the last big number er and not on the
in the mirror she was shocked. She'd lost	weight	she was dressed abominably her hair looked
learning difficulties [F01] I've lost a lot of	weight.	[tc text=laughs] [M01] Have you [F01] Erm I'

Figure 2-3 Concordances of 'weight' from the Collins WordbanksOnline English Corpus

Though *data-driven learning* has been widely accepted among those who exploit corpus resources for pedagogical purposes, the term itself seems misleading: DDL is a pedagogical approach rather than learning *per se*. Central to Johns' (1991) argument is that DDL departs from the traditional rule-based, deductive pedagogical approach in which the teacher explicitly presents prescribed language rules/patterns from textbooks, and constructs utterances by applying these rules/patterns and filling lexical items into the slots in the patterns. DDL, on the other hand, provides the learner with the opportunity to explore a wealth of language in use and thereby induce patterns from the bottom up, instead of being fed the prescribed language rules/patterns. In a DDL setting, the learner is encouraged to assume an active role in interrogating corpora to explore how a particular word is typically used in the language. After a thorough examination of language data in corpora, the learner becomes more aware of the typical behaviour of the query word, and is thus able to come up with his/her own generalization of its collocational patterns. Whereas the



*rule-based, deductive* pedagogical approach aims to encapsulate linguistic ‘competence’, the *data-driven, inductive* one provides access to linguistic ‘performance’ (Johns, *ibid.*). From the above description of what constitutes DDL, it is clear that the distinction between the *rule-based, deductive* approach and the *data-driven, inductive* approach seems to be a pedagogical one, rather than an acquisitional one, as these approaches refer to the ways in which the information to be learned is delivered to the learner, imparted by the teacher or induced by the learner. In this sense, Johns’ (*ibid.*) DDL has only gone so far as to advocate the learner’s self-directed induction of language patterns, but does not further explain how *learning* actually takes place through such an inductive approach. Therefore, the term *data-driven learning* may misrepresent what is actually entailed in such endeavours, as it does not acknowledge the fundamental role of *pedagogical mediation*, and oversimplifies the highly complex mechanism of *learning* taking place as a result of such mediation (the notion of *mediation* and the resulted learning is addressed at length in 2.4.2). This study therefore uses the term *data-driven approach to learning collocations* (DALC) to refer to the pedagogical approach in which the learner accesses and explores language in use in corpora to aid collocation learning.

McEnery and Wilson (1997:6) characterize corpus-assisted language learning as:

- Discovery learning – learners discover the features about language through self-access exploration of corpus data.
- Mediated learning – the corpus is not a source of didactic learning: rather, it is a medium through which learning may be achieved. Students learn through the process of interacting in some way with the corpus (e.g. via practical grammatical analysis), rather than from its explicit content. (Contrast the traditional textbook, where the students do learn directly from its contents.)
- Divergent learning – learners do not follow the same path through the data and do not necessarily generate the same finding.
- Directed learning – learners are directed by the teacher but led by themselves through the corpus consultation.

McEnery and Wilson’s (*ibid.*) characterization aptly encapsulates the essence of corpus-assisted language pedagogy: *discovery learning and mediated learning* depict the *learning processes*, whereas *divergent learning* concerns the *learning outcomes*; *directed learning* points out the different roles played by the learner and the teacher in such a learning setting. The following subsections provide the rationale underpinning DALC, incorporating notions such as authenticity, abundant language data and autonomous learning.



### 2.3.2.1 Authenticity versus genuineness

Corpus-assisted language pedagogy is characterized by the use of language data drawn from written and/or spoken texts that originally serve communication purposes in real-life contexts, as opposed to classroom use (Aston, 1995; Braun, 2005; Cobb, 1997; Flowerdew, 2001; Sinclair, 2004; Wichmann et al., 1997). As Tsui (2004:40) aptly puts it, corpora provide a better understanding of 'how language is actually used rather than how language is perceived to be used'. Braun (2005) commends corpus data for being:

- realistic, showing language in real use;
- rich, providing more (and more diversified) information than dictionaries or reference grammars can;
- illustrative, providing actual patterns of use instead of abstract explanations;
- up-to-date, revealing trends in language use and evidence for short-term historical change.

A perpetual challenge facing FL learning is limited naturalistic exposure to the TL, compounded by the potential pitfall that 'concocted' textbook examples (Carter and McCarthy, 1995:154) might not truthfully reflect how the language is used in real-life contexts. In this regard, corpora can be a particularly valuable resource for providing examples of target language in use.

Notwithstanding that many researchers see real language data as the major advantage a corpus has to offer, few studies have explored the learner's perceptions thereof, with the exception of Chambers (2005) and Farr (2008). Chambers (ibid.) examined 14 language learners' processes of corpus consultation in terms of analytical skills, problems encountered and learner evaluation of the activity. Despite the small number of participants, there has been a general consensus as to the perceived authenticity of corpus language which has been described as 'authentic', 'real', 'up-to-date' and 'relevant', as opposed to the invented examples in textbooks. Similarly, Farr (ibid.) found that student teachers perceived the major benefit of corpus-assisted language pedagogy as providing direct access to real language use.

The fact that the language data in corpora originally serves real-life communication purposes, as opposed to pedagogical ones, has led many to believe that corpus language is intrinsically authentic. However, some scholars problematize the view that authenticity is an inherent quality in texts. Widdowson (1978), among others, argues

that although the origin of corpus language is genuine, the format in which corpus data is presented (i.e., concordance lines) inevitably denudes the language and its original contexts on which meaning depends. The 'complex interplay of linguistic and contextual factors whereby discourse is enacted' (Widdowson, 2000:7) underlying authentic language is therefore lost in the conversion into concordances. Hence, Widdowson (1978:80) draws a distinction between *genuineness* and *authenticity*:

[g]enuineness is a characteristic of the passage itself and is an absolute quality. Authenticity is a characteristic of the relationship between the passage and the reader and it has to do with appropriate response.

Widdowson (2000:7) goes on to argue that *authenticity* can only be realized through the *recontextualization* of corpus language:

[t]he texts which are collected in a corpus have a reflected reality: they are only real because of the presupposed reality of the discourses of which they are a trace. This is decontextualized language, which is why it is only partially real. If the language is to be realized as use, it has to be recontextualized. The textual findings of frequencies and co-occurrences have to be contextually reconstituted in the classroom for their reality to be realized, and this reconstitution must obviously be based on very different contextual conditions than those which activated the texts in the first place.

In this sense, pedagogical mediation is necessary to contextually reconstitute the linguistic descriptions derived from corpus findings, so that the learner may be able to relate to their personal experience. Following Widdowson (1978), Van Lier (2000:128) contends that '[a]uthenticity is the result of acts of authentication, by students and their teacher, of the learning process and the language used in it [...] authentication is basically a personal process of engagement'. In other words, attested instances of language in corpora need to be mediated by pedagogical actions, in order to allow *authentication* on the part of the learner. Wray (2000:484) echoes the call for pedagogical mediation when exposing the learner to real language data, '[t]he more natural the data that are being presented to the learner, the more they need to be subjected to control and guidance in delivery'. It is precisely the *genuineness* of corpus language that necessitates contextually appropriate scaffolding. Corpora promise to provide what traditional FL pedagogy fails to offer, namely genuine TL, provided that adequate pedagogical mediation is carried out. Whereas corpus language may not be *authentic* by nature, it nonetheless offers raw data of which learners can be guided to avail themselves. For example, Bernardini (2002) guided student translators to compare the frequencies of occurrences of a lexical collocation

across genres and registers in various corpora, in order to illustrate the differences in style, underlying ideology, recurrent themes, and so on. As the success of translation depends on the appropriateness of the language forms used in a particular genre/register, the corpus-driven mediation to demonstrate linguistic tendencies was highly pertinent to student translators' career needs. The activity thus led them to perceive corpus consultation as relevant to their current and/or future profession. Aston (1995:263) underscores the importance of the perceived relevance to the learner:

[c]orpus-based methods have proved to be a powerful means of highlighting patterns of repetition and variation in text. Whether these methods can be of value for learning depends on the extent to which the patterns highlighted are relevant to the learner, emerging in contexts which facilitate the restructuring of knowledge, and in such a manner as to make this knowledge available for use.

It is clear from the discussion above that what is important is not *what* corpora have to offer, but *how* they can be made useful and relevant to the learner through his/her authentication induced by pedagogical mediation (see 3.4 for how the DALC task was designed to induce authentication and mediate collocation learning).

### 2.3.2.2 Profusion

Another major strength of corpora lies in the amount of language data. Modern mega-size computerized corpora store tens or even hundreds of millions of words. The wealth of language data in corpora makes linguistic regularities salient. As Tribble and Jones (1990:3) put it, '[t]he concordancer lets you rearrange texts in such a way that it becomes possible to see patterns that would not otherwise be visible'. Because of their potential to highlight linguistic regularities, corpora have been used for language description and analysis such as lexico-grammatical patterning based on co-occurrence of linguistic items, discourse analysis, and genre analysis by examining the co-occurrence of groups of linguistic items and processes (Tsui, 2004).

The frequency of occurrence is one of the key determinants of *noticing*, a necessary and sufficient condition for converting *input* to *intake*, hence *learning* (Schmidt, 1990). In this sense, the more frequently a collocation occurs, the more likely it is to be *noticed* by the learner. Furthermore, a body of empirical evidence suggests that a learner needs to encounter a word 5 to 16 times or more before s/he acquires it (Nation, 1990). It is therefore reasonable to expect a learner to need at least as many



times to form collocational links. However, in FL contexts where naturalistic exposure is lacking, it may take years for the learner to incidentally encounter the same collocation a number of times. In such cases, corpus consultation can be an efficient alternative to provide intensive exposure to a wealth of real language from a range of contexts which would take years for a FL learner to encounter incidentally. Mega-size corpora provide easy access for the learner to be exposed to abundant instances of how a target word is typically used, in the volume that neither naturalistic exposure nor classroom instruction can rival. As Thurstun and Candlin (1998:270) state, corpora provide 'the opportunity to condense and intensify the process of learning through exposure to multiple examples of the same vocabulary item in context, and to promote awareness of collocational relationships'. It is precisely through this condensation and intensification of exposure that the collocational patterns and contextual preferences of a word are made salient for the learner (Aston, 1995; Barlow, 1996). Such intensive exposure to multiple instances in which a target word is embedded helps to *consolidate* and *enhance* knowledge of the particular word. If the learner already has a particular lexical pattern in mind, encountering the same pattern repeatedly in multiple contexts in a corpus is conducive to *consolidating* his/her knowledge of the pattern. The profusion of corpus data provides rich resources for 'recycling' previously studied words/patterns in a principled way. 'Recycling' is of paramount importance for consolidating vocabulary knowledge (Schmitt, 2008). Furthermore, in the course of corpus exploration, the learner may incidentally pick up a grammatical usage or a collocation that has not been learned before, hence *enhancing* knowledge of word use.

The profusion of language data in corpora not only makes collocational patterns salient, but also has the potential to raise the learner's awareness of the contextual preference and frequency distribution of a word. For instance, it is clear from the concordances below that the typical grammatical pattern of the query word *tantamount* is *be tantamount to*. A closer look at the objects following the pattern *be tantamount to* reveals that it typically precedes things with a negative meaning sense. In this case, the collection of texts in which the query word is embedded illustrates the lexical behaviour as well as the semantic prosody of the word. Also, through an observation of multiple contexts in which the word *tantamount* is embedded, the learner may become aware that the word tends to occur more frequently in formal



registers. In general, corpora provide a rich source of information on *word use* (e.g., grammatical function, collocation and constraints on use such as register or frequency).

<p>..of the real estate business is the idea that renting a home is and collecting. To outlaw them because of how they look is ...make external beauty the sole criterion for the beautiful is ...Logging on before becoming a modern-day Red Baron is .....that "taking up one's cross" is, for the Christian scholar, detriment of sign language could be seen through this lens as .....other test and its warning that U.N. sanctions would be</p>	<p>tantamount to throwing money away. It's a useful fiction for tantamount to banning cars because of their hood ornament. tantamount to penalizing the physically unattractive for a tantamount to suicide, however, as human opponents are far tantamount to sacrificing one's passion for absolute and tantamount to "genocide," a term that was not shied away tantamount to a declaration of war. Those statements have</p>
--	--

**Figure 2-4** *Concordances of 'tantamount' from the Corpus of Contemporary American English*

Empirical evidence also demonstrates that the wealth of corpus data raises language awareness. O'Keeffe and Farr (2003) illustrated with examples that corpora can be used to raise awareness of the following aspects of language: lexical patterns, grammatical patterns, register-specific linguistic choices and sociocultural grammatical choices. In similar vein, Sealey and Thompson (2007) found that corpus exploration prompted metalinguistic discourse among young L1 learners, a process in which they demonstrated and negotiated their metalinguistic awareness. In the light of participant reflection, Chambers (2005) found that after corpus consultation, the participants were sensitized to the 'uncertainty' of language and became dubious about the authority of prescriptive grammars.

Although the profusion of corpus data provides insights into language regularities in such a way that is otherwise unlikely, the amount can be a cause for concern when it comes to learner self-access corpus consultation. A mega-size corpus can easily generate hundreds of concordance lines for frequent words, the amount of which may be overwhelming for novice users (Braun, 2005). Meunier (2002) further points out that corpus results can be messy, ambiguous or misleading. In this regard, Kennedy and Miceli (2010:41), drawing on the findings of their case studies, suggest that one of the principles underpinning effective corpus consultation is that 'working with corpora requires greater preparedness to proceed by trial and error than work with other reference resources, and acceptance of the uncertainty of finding a satisfactory answer'. A sense of *uncertainty* is arguably what the learner may experience in real-life communication contexts outside the language classroom, as opposed to 'living in the false expectation that clear and simple rules can always be devised' (Boulton, 2009:39). The awareness and acceptance of *uncertainty* is therefore not only important in corpus consultation but also language learning in general. As regards

vocabulary learning with corpora, while acknowledging the problems that large corpora may pose, Cobb (1999) is less concerned about the *messiness* of corpus data, as he maintains that words can be seen in multiple contexts rather than in isolation and that learners would not be distracted by the flow of discourse. To scaffold learners to fare with corpus data more successfully, it is therefore important to take cognizance of the word of caution above when designing and implementing corpus-assisted language tasks. In summary, the present study aims to facilitate collocation learning, so the potential that corpora have in highlighting recurring patterns can be usefully exploited to serve the pedagogical purpose here. On the other hand, the potential pitfall of *messiness* accompanying a wealth of language data necessitates careful scaffolding (e.g., via training on concordancing skills and awareness-raising) to better prepare learners to make the full use of corpus resources for language learning.

### **2.3.2.3 Learner autonomy**

Corpus-assisted language pedagogy provides the learner with an opportunity to develop learner autonomy by exploring and analyzing real language data and eventually coming to their own generalizations about language regularities (Boulton, 2009; Chambers and O'Sullivan, 2004; Gable, 2001; Gavioli and Aston, 2001; Kaltenbock and Mehlmauer-Larcher, 2005; O'Sullivan, 2007). In a traditional teacher-centered EFL classroom, the teacher is the imparter of linguistic knowledge and the student is viewed as the passive recipient of knowledge. In contrast, a data-driven approach to language learning induces learner-centered discovery learning in which linguistic regularities and/or variations are to be discovered by the learner, rather than transmitted by the teacher. Such inductive pedagogical approaches are not new to SL/FL pedagogy, however, the induction mediated by an interaction with a repository of real language data is what distinguishes corpus-assisted language pedagogy from other inductive pedagogical approaches. Johns (1991) observes that the data-driven pedagogical approach 'stimulates enquiry and speculation on the part of the learner, and helping the learner also to develop the ability to see patterning in the target language and to form generalizations to account for that patterning'. It requires the learner to assume the role of a researcher instead of a passive recipient of knowledge imparted from the teacher. Student researchers need to assume the responsibility of exploring language in use, analyzing the data, observing the usage

and eventually generalizing the patterns of how a given word is typically used. The teacher, on the other hand, acts as a research facilitator or director in such a learner-centered inquiry, rather than the transmitter of knowledge (Flowerdew, 1996). Recognizing the changes in the classroom dynamic in corpus-assisted learning settings, Bernardini (2004) claims that the teacher here acts as a *learning expert* instead of a *language expert*, as s/he facilitates the learner's linguistic exploration, but does not prescribe what to be learned. While it is generally agreed that the data-driven pedagogical approach is conducive to developing learner autonomy, Bernardini (ibid.) further contends that it is a democratization of learning settings, as it empowers the learner as well as the teacher. The discovery nature of the data-driven pedagogical approach challenges learners linguistically and cognitively, but also motivates them to inquire and explore. In advocating the pedagogical use of corpora in language classrooms, Leech and Candlin (1986:xvi) call for:

classroom access to language databases, lexicographic and grammatical corpora, oriented to learners' interlanguages and displayed in terms that learners (not only lexicographers and grammarians) can understand... What of the process tasks to be made available as problem-solving exercises for learners to apply to text? Two characteristics will predominate: they will have to involve learners in solving problems and experimenting with language learning, and they will need to be differentiated in terms of offering alternative routes, varying levels of demand and attainment, and alternative possibilities of solution. In short, they will need to mirror the cognitive requirements of language learning.

Although there has not been any work that specifically tests the claim that corpus-assisted pedagogical approaches promote learner autonomy, two studies, Bernardini (2002) and Yoon and Hirvela (2004), nonetheless reported learner perceptions of increased confidence arising from self-access corpus use. Bernardini's (ibid:179) participants liked 'the idea of feeling competent, of having a say in what was happening around them', while Yoon and Hirvela's (ibid.) felt confident using corpora as a tool for testing linguistic hypotheses. Perceived confidence or learner empowerment is arguably a key element in sustaining motivation, and may therefore go a long way towards fostering learner autonomy.

It is worth noting that the learner autonomy advocated here does not mean wholesale abandonment of teacher instruction or pedagogical actions, on the contrary, only appropriate pedagogical mediation which is fine-tuned to the learner's needs in combination with learner-initiated corpus exploration can lead to authentication of corpus language on the part of the learner and facilitate autonomous discovery



learning.

## **2.4 How collocation learning occurs with DALC – Theoretical underpinnings from cognitive and sociocultural perspectives**

DALC is a relatively new pedagogical approach, having emerged with the advent of computer and corpus technology instead of stemming from any particular SLA theory. It does not fit neatly into any single theoretical framework of SLA, hence, I argue that *learning* with DALC can only be accounted for by a synthesis of *cognitive approaches to SLA* and *sociocultural theory* (SCT). This section reviews *cognitive approaches* (2.4.1) and SCT (2.4.2) in the SLA literature. *Connectionism* accounts for how *collocational knowledge* is mentally represented in the mind (2.4.1.1). Also, the emphasis on *frequency* at the heart of connectionist models provides a basis for the pedagogical use of concordances. *Information processing approaches* (2.4.1.2) such as *language learning strategies* account for the cognitive processes whereby the learner encodes corpus data as a prelude to actual *learning*. The *depth of processing hypothesis* and, more recently and specifically, the *involvement load hypothesis*, provide explanations on the conditions for *learning* to take place. Overall, the notion of *mediation* within SCT underpins all types of technology-enhanced learning, and DALC is no exception. This section finally considers how these theoretical frameworks reconcile to account for *collocation learning* that takes place as a result of DALC (2.4.3).

### **2.4.1 Cognitive approaches to SLA**

Cognitive approaches to SLA, such as *connectionism* and *information processing approaches*, have important implications for how language learning takes place and how language knowledge is represented within the mind. Informed by *connectionism*, I conceptualize the mental representation of *word knowledge* as networks of connections, of which collocational connection is one. Learning in this view occurs as a connection strengthens with repeated exposure to the target language point. While such an associative mechanism only accounts for part of the picture of language learning, *information processing* explains the interaction between the learner and language input as the prelude to *learning*. The following subsections address



*collocation learning* from various positions within the broad spectrum of cognitive approaches to SLA.

#### **2.4.1.1 Learning as establishing connections - Connectionism**

As noted in Section 2.1, generative accounts of language, particularly UG, have dominated the fields of first language acquisition (L1A) and SLA (at a later date) in the last half-century. Generative accounts argue for a universal grammar intrinsic in the human mind and a variable lexicon that needs to be acquired extrinsically. Language in this view is mentally represented as a rule-based system and language acquisition is thus rule-governed. However, in the past two decades or so, there has been growing discontent with the fact that generative approaches fail to provide an adequate account of how SLA occurs. Many turn to connectionist/emergentist models that contrast strikingly with generative accounts in that they conceptualize the mental representation of language as *exemplar-based*, as opposed to the *rule-based* view held by generative accounts.

##### **1) The mental representation of collocational knowledge**

Connectionist/emergentist models hold that what constitutes language knowledge is not symbolic representations (i.e., abstract grammar rules) as argued by generative accounts, but a vast store of previously experienced utterances, or *exemplars* (N. Ellis, 1998). These exemplars are interconnected to form a complex system of networks. Having the roots in the field of psychology, connectionist models regard language representation as neural networks that are composed of information nodes connected by pathways (Gass and Selinker, 2008). Language learning in this view is the processes of establishing and strengthening the connections within neural networks through activation or use. Word knowledge, in this account, is conceptualized as one of the neural networks that has complex clusters of connections such as form (written and spoken), meaning, collocation, association, and so on. Also, words are no longer viewed as discrete units to be learned or taught independently of links to other words. More specifically, *collocational knowledge* is one of the information nodes that are connected to a target word and to other links of the word. Note that within the connectionist framework, exemplars are connected such that they reflect *rule-like* behaviour, but it does not follow that language representation and acquisition are

*rule-based*. Language rules/patterns do not exist *a priori* in the learner's mind, but emerge as a result of repeated exposure to exemplars in language input. As Randall (2007:21) puts it, '[l]anguage rules "emerge" from the input as a series of probabilities of the co-occurrence of certain features not as symbolic representations such as grammar rules'. In opposition to generative view of language, N. Ellis (1996:364) writes, 'rule-like behaviour does not imply rule-governed behaviour'.

Though connectionists hold that language acquisition and representation are exemplar-based, it does not follow that language production will reproduce exactly the same exemplars that were experienced previously. Rather, language production operates along a formulaic-creative continuum (Bolinger, 1976; Bolinger and Sears, 1981; Nattinger and DeCarrico, 1992; Pawley and Syder, 1983; Weinert, 1995; Wray, 2002), which coincides closely with Sinclair's (1991) *open-choice principle* and *idiom principle* (see 2.1). The metaphor of a continuum is particularly useful here, because it takes an eclectic stance to encompass the creativity as well as the formulaicity in language production. Not all utterances, be it spoken or written, are assembled from scratch every time one wishes to communicate (i.e., on the creativity end or open-choice principle), but neither are they retrieved whole from the mind without any degree of analysis (i.e., on the formulaicity end or idiom principle). The parallel processes of creative construction and formulaic language use give a more accurate account of language production than either side of the analytic-synthetic dichotomy.

## **2) Frequency as a determinant of forming collocational connections**

As regards language learning, while innatists hold that there is a neurological module specific to language learning (e.g., UG), connectionists/emergentists argue that language learning as well as other higher-order mental activities take place in a generic learning mechanism within an individual. Learning, in this view, occurs when the generic learning mechanism interacts with the environment. More specifically, language learning takes place as repeated exposure to language input establishes or strengthens connections to the existing neural networks in the mind. As N. Ellis (2002:144) argues:

[l]anguage learning is the associative learning of representations that reflect the probabilities of occurrence of form-function mappings. Frequency is thus a key determinant of acquisition because 'rules' of language, at all levels of analysis (from phonology, through syntax, to discourse), are

structural regularities that emerge from learners' lifetime analysis of the distributional characteristics of the language input.

Following N. Ellis (ibid.), Randall (2007) maintains that the interconnectivity in the brain mirrors the probabilistic relations between features in the language. The strengths of connections vary depending on the frequency with which a particular linguistic feature occurs and the frequency with which the learner encounters this feature in language input (Lightbown and Spada, 2006). N. Ellis (2002:166-167) underscores the role of frequency in language acquisition:

[f]or language learners to be accurate and fluent in their generalizations they need to have processed sufficient exemplars that their accidental and finite experience is truly representative of the total population of language of the speech community in terms of its overall content, the relative frequencies of that content, and the mappings of form to functional interpretation.

Within the connectionist framework, learning collocations consists of forming or strengthening the connections between the node word and its collocates through repeated activation or use. As noted in Section 2.2.2.2, collocations are usage-based and language-specific, so they do not always follow semantic reasoning (e.g., *sour cream* but not *\*rancid cream*) and cannot be literally transferred from L1. Therefore, *exemplars* are particularly important in learning such an arbitrary aspect of language as collocation. However, in many FL contexts where repeated naturalistic exposure is insufficient, it seems unrealistic to expect collocation learning to take place through incidental encounters over time. In this case, a wealth of language data in corpora can be usefully exploited to provide intensive exposure to exemplars necessary for strengthening collocational links, hence the corpus-assisted, data-driven approach to learning collocations.

Whereas this study embraces a broadly connectionist account of the mental representation and acquisition of collocational knowledge, it did not employ the research methodology conventional to connectionists, namely computer simulation (e.g., Sokolik and Smith's (1992) investigation on the assignment of gender to French nouns; N. Ellis and Schmidt's (1997) use of an artificial language in a laboratory situation looking at the adult acquisition of plural morphology). Computer simulation can only process small, well-contained samples (Mitchell and Myles, 2004). Collocational relations, though not as open-ended as generative accounts claim, are far more complex than closed sets of connections such as subject-verb agreement or gender assignment to nouns. In other words, connections such as subject-verb



agreement or gender assignment to nouns are relatively straightforward in that they are based solely on grammaticality, while the choice of collocates may depend on a range of factors such as usage, context or style, the interconnections of which are well beyond the control of computer modeling. Acknowledging the limitation of computer simulation on modeling collocation learning, the present study accords with the connectionist approach to SLA only insofar as it accounts for the mental representation and associative learning mechanism of collocational knowledge.

In sum, connectionist models provide explanations of how language learning occurs as a result of repeated activation of connections that are embedded in complex neural networks within the mind. Learning is thus a dynamic associative process rather than a static phenomenon, entailing the strengthening as well as attrition of connections.

#### **2.4.1.2 Learning as processing information – Information processing approaches**

The previous section has discussed the connectionist account of how collocational knowledge is acquired through an associative mechanism and how it is mentally represented in the mind. In this section, still within the broad spectrum of cognitive approaches, I narrow down the focus to the *cognitive processes* leading to actual learning as the learner engages in a corpus-assisted language task. *Information processing approaches* are concerned with how human mental capacities process information input and how these mechanisms work within the context of learning (Gass and Selinker, 2008). A useful metaphor here is human brain as a computer receiving and encoding input in working memory, converting *input* to *intake* through *noticing*, committing to long-term memory and decoding for output as necessary. The present study aims to explore the ways in which the learner processes language input in the format of concordances as s/he consults corpora, so the *cognitive strategies* manipulating the input and the *metacognitive strategies* orchestrating the encoding processes are the foci of this investigation. The *depth of processing hypothesis* and the *involvement load hypothesis* provide a sound basis for inductive pedagogical approaches such as DALC.



1) Language learning strategies for processing information

Learner strategies fall into two broad categories, *strategies for learning* and *strategies for use* (Cohen, 1998): this section only reviews the former as *learning* is the focus of the study. Recent research on language learning strategies (henceforth LLSs) has its roots in the ‘good language learner’ studies in the 1970s (e.g., Naiman et al., 1978; Rubin, 1975; Stern, 1975). Since then, much has been written regarding the conceptual and classificatory issues in this area. Having reviewed a body of research on LLSs, Gass and Selinker (2008:440) highlight the core concept:

potential improvements in language learning related to the selection of information from the input and the organization and integration of it in terms of learner systems. The ways in which information is selected from the input are an important part of the concept.

The table below summarizes some of the definitions of LLSs in the literature.

Table 2-4 Definitions of language learning strategies

Author	Definition
Bialystok (1978:71)	optional means for exploiting available information to improve competence in a second language.
Wenden and Rubin (1987:19)	any sets of operations, steps, plans, routines used by the learner to facilitate the obtaining, storage, retrieval, and use of information.
Stern (1992:261)	the concept of learning strategy is dependent on the assumption that learners consciously engage in activities to achieve certain goals and learning strategies can be regarded as broadly conceived intentional directions and learning techniques.
Oxford (1999:518)	specific actions, behaviours, steps, or techniques that students use to improve their own progress in developing skills in a second or foreign language.
Cohen (1998:4)	those processes which are consciously selected by learners and which may result in action taken to enhance the learning or use of a second or foreign language, through the storage, retention, recall, and application of information about that language.

Given the strong emphasis on the manipulation, storage and retrieval of information, it is clear that LLS is grounded in *information processing approaches*. In their seminal work on learning strategies in SLA, O'Malley and Chamot (1990:1) explicitly align themselves with ‘a cognitive information processing view of human thought and action’, defining learning strategies as ‘special ways of processing information that enhance comprehension, learning, or retention of the information’. Notably, in considering the functions of learning strategies, Macaro (2006) argues that these strategies only occur in working memory, and become other constructs elsewhere. In this sense, LLSs represent the mental functioning in working/short-term memory that process language input before it is ready to be committed to long-term memory for

storage and future retrieval. This micro, information processing view is highly pertinent to the present study, which aims to uncover the mental processes that take place as the learner interacts with corpus data as a means of inducing collocations.

In addition to delineating the construct of LLSs, considerable effort has been devoted to classify such strategies (e.g., Bialystok, 1978; Naiman et al., 1978; O'Malley and Chamot, 1990; Oxford, 1990; Rubin, 1981; Stern, 1992). In particular, the taxonomy of LLSs developed by O'Malley and Chamot (1990) fits well with the present study on account of its strong emphasis on cognition (in its broad sense). Firmly grounded in cognitive theory of learning and a body of empirical evidence, O'Malley and Chamot (ibid.) were able to develop a comprehensive taxonomy that systematically encapsulated a vast array of LLSs used by L2 learners. O'Malley and Chamot (ibid. 44-45) identified three broad categories of LLSs: 1) *cognitive strategies*, which 'operate directly on incoming information, manipulating it in ways that enhance learning'; 2) *metacognitive strategies*, which are associated with 'higher order executive skills that may entail planning for, monitoring, or evaluating the success of a learning activity'; and 3) *socioaffective strategies*, which 'represent a broad grouping that involves either interaction with another person or ideational control over affect'. The taxonomy is summarized in the table below:

**Table 2-5 Taxonomy of LLSs in O'Malley and Chamot (1990)**

<b>Cognitive strategies</b>	repetition, resourcing, grouping, note taking, deduction/induction, substitution, elaboration, summarization, translation, transfer and inferencing.
<b>Metacognitive strategies</b>	planning, directed attention, selective attention, self-management, self-monitoring, problem identification and self-evaluation.
<b>Socioaffective strategies</b>	questioning for clarification, cooperation, self-talk and self-reinforcement.

Few studies have examined the thinking processes that underlie corpus consultation behaviour, with the exception of Sun (2003), which will be discussed in Section 2.5.2. The present study uses O'Malley and Chamot's (ibid.) taxonomy of LLSs as a preliminary framework to unpack the processes. This taxonomy is particularly useful here because the mental processes mediated by corpus consultation are highly specific and under-researched, a generic and comprehensive framework such as this one would serve as a reasonable starting point for my exploratory endeavour to understand the cognitive processes that underlie corpus consultation.

## 2) Depth of processing

In the field of cognitive psychology, Craik and Lockhart (1972) propose the *depth of processing hypothesis* in their pioneering work, arguing that the likelihood of a piece of information being committed to long-term memory is dependent on the depth with which it is initially processed in short-term/working memory. All other factors being equal, a higher degree of information processing is more likely to lead to *learning* than a lower degree one. A large body of empirical evidence has since then corroborated the fact that cognitively demanding activities which require a more intense manipulation of information input yield more effective learning (cf. Baddeley, 1997; Craik and Tulving, 1975; Zechmeister and Nyberg, 1982).

While the *depth of processing hypothesis* has been widely tested in the field of cognitive psychology, it is not until recently that attempts have been made to use this notion to specifically account for vocabulary learning in SLA. Laufer and Hulstijn (2001) propose the *involvement load hypothesis* based on an extensive review of empirical studies looking into the effectiveness of various vocabulary tasks, re-examining the findings in the light of learner involvement load induced by task requirements. The *involvement load hypothesis* consists of three assumptions:

- 1) retention of words is conditional upon the learner's involvement in the task, which is composed of three factors: *need*, *search* and *evaluation*;
- 2) words processed with a higher involvement load will be retained better than those processed with a lower involvement load;
- 3) tasks with a higher involvement load will better facilitate vocabulary retention than those with a lower involvement load.

In terms of the learner's involvement, the *need* component is the drive to comply with the externally imposed or self-imposed task requirements; the *search* component is 'the attempt to find the meaning of an unknown L2 word or trying to find the L2 word form expressing a concept (e.g., trying to find a L2 translation of a L1 word) by consulting a dictionary or another authority (e.g., a teacher)'; and the *evaluation* component 'entails a comparison of a given word with other words, a specific meaning of a word with its other meanings, or combining the word with other words in order to assess whether a word (i.e., form-meaning pair) does or does not fit its context' (ibid:14). If the hypothesis bears out, the vocabulary tasks designed to induce higher levels of learner involvement (as measured by the *need*, *search* and *evaluation* components) will better facilitate vocabulary learning. In the present study, DALC is



an inductive pedagogical approach (see 2.3.2.3) which necessitates a *search* and *evaluation* of collocations, as the learner is required to generalize collocational patterns from a host of real language data in corpora. Unlike other types of reference resources (e.g., dictionaries or grammar books), corpora do not present straightforward linguistic descriptions generalized by lexicographers or grammarians, they provide raw materials from which the user has to induce the target linguistic features. Through such data-driven, induction processes, the learner engages more deeply in encoding the language input from concordances than simply being fed the target collocations for rote memorization. It is precisely such a cognitively intense manipulation of corpus data required on the part of the learner that makes DALC conducive to *learning*. Hence, the *depth of processing hypothesis* in cognitive psychology and the *involvement load hypothesis* in SLA underpin the use of corpus-driven, inductive language tasks to facilitate collocation learning (see the DALC task in 3.4).

## **2.4.2 Sociocultural theory**

Whereas cognitive approaches to SLA explain how the learner employs mental capacities to process language input in working memory and how language knowledge is represented in long-term memory, such approaches do not seem to attach as much importance to the mediating role of input as SCT does. Language input in cognitive approaches is treated as a passive object to be processed, but SCT highlights the active role of language input as a mediating means which synchronizes an individual and the social and material world around him/her (Block, 2003). Hence, SCT provides a sound theoretical basis for exploiting technology to mediate language learning, namely technology-enhanced learning (TEL) or computer-assisted language learning (CALL). How learning occurs through *mediation* within the framework of SCT is elaborated in the following subsections.

### **2.4.2.1 Mediation and appropriation**

As the name suggests, SCT sees the social context at the heart of the development of mental processes, but it does not necessarily reject the essential role of biologically endowed cognitive mechanisms (Neisser, 1992). As noted in Section 2.4.1.2, cognitive approaches to SLA focus on the computational processes in which the brain



receives input from the external world, processes and stores information, and retrieves to produce output in response to situational demands. Within the sociocultural framework, however, biologically endowed capacities are necessary but insufficient conditions for human development, all higher forms of mental activities are *mediated* by culturally constructed auxiliary means. Therefore, the concept of *mediation* is key to learning in SCT (Lantolf and Appel, 1994). *Mediation* is defined by Lantolf and Thorne (2006:79) as:

the process through which humans deploy culturally constructed artefacts, concepts, and activities to regulate (i.e. gain voluntary control over and transform) the material world or their own and each other's social and mental activity.

It is through *mediation* that the control of mental capacities shifts from the environment to the individual, which leads to voluntary regulation over mental functioning (Wertsch, 1998). To put it differently, *mediation* enables the cultural/interpersonal plane to interact with the biological/intrapersonal plane so as to give rise to the development of higher-order mental functions such as *learning* (Lantolf, 2000). Learning, from this perspective, takes place as an individual uses socially constructed mediating means to gain control over his/her own mental activities (Lantolf and Thorne, 2007; Wertsch, 1991). This process is called *appropriation*, a process whereby learners 'take over (or *appropriate*) new knowledge or skills into their own individual consciousness; successful learning involves a shift from inter-mental activity to intra-mental activity' (Mitchell and Myles, 2004:145). Drawing on the concepts of *mediation* and *appropriation* within SCT, learning collocations in a FL is a process whereby the learner appropriates collocations recurring in the TL (i.e., the interpersonal plane) to restructure their mental lexicon (i.e., the intrapersonal plane) through the use of mediating means such as keeping notes or consulting a dictionary. More specifically, DALC entails the process in which the learner explores corpus data (i.e., mediation) to induce collocational patterns to be integrated into his/her existing word knowledge (i.e., appropriation). It is noteworthy that *appropriation* is not merely the passing of the external to the internal, rather, it is the interweaving of the external and the internal that creates 'a synthesized new state' (Block, 2003:103).

### 2.4.2.2 Tools for mediation

Since the concept of *mediation* is central to SCT, the tools for mediation merit discussion. Tools for mediation can be either *physical* or *symbolic* (or 'psychological' in Vygotsky's term). *Physical tools* such as hammers or screwdrivers extend the reach and power of an individual's physical strengths and enable him/her to act on objects and shape the course of events. Whereas *physical tools* are directed outwardly to alter the material world in which the tool user lives, *symbolic tools* (e.g., maps, street signs or language) can be directed outwardly to influence external objects or another individual or inwardly to regulate his/her own mental activities (Vygotsky, 1997). *Symbolic tools* may not affect the material world in the ways physical tools do, but they nonetheless are able to 'radically reconstruct the whole mental operation' of oneself and of others (ibid.:63). Bruner (1966) regards *symbolic tools* as 'cultural amplifiers', arguing that just as *physical tools* enhance an individual's biological strengths (e.g., a hammer increases one's strengths), *symbolic tools* amplify cognitive abilities such as memory, attention or rational thinking. Notably, *symbolic mediation* not only brings about 'quantitative improvements in terms of speed and efficiency' but also 'a qualitative transformation' in the thinking processes (Wertsch, 2007:17). As Vygotsky (1981:137) states, 'by being included in the process of behaviour, the psychological tool alters the entire flow and structure of mental functions'. Hence, *symbolic mediation* enables *quantitative* and *qualitative* appropriation to take place and leads to the development of higher-order mental functions. To put theory into practice, it is hoped that the *symbolic mediation* of corpus consultation will 1) *quantitatively* expand the learner's repertoire of collocations; 2) effect a *qualitative* transformation whereby the learner becomes sensitized to the ubiquity of formulaic sequences such as collocation in the language.

The most powerful symbolic tool of all, according to Vygotsky (1978), is *language*. Language can be used as a means to communicate as well as to organize and manage mental functioning. Within sociocultural SLA, as R. Ellis (Ellis, 2003) points out, 'language learning involves both developing the means for mediating learning, i.e. the tools, and the language itself, i.e. the object'. Language, in this sense, has a 'double character' because it is an 'object in the world around us that we can reflect on, and they mediate our interaction with the world' (Bodker, 1997:150). The double character

of language is nowhere more evident than in corpus-driven pedagogy: on the one hand, language regularities emerging from corpus data are the object for such a learning activity; on the other hand, the human mind does not simply mirror what is presented in the concordances, but processes the language input (i.e., corpus data) as a mediating tool to restructure the mental lexicon, hence *learning*. Language data in corpora serves as a means and an end at the same time.

Taken together, DALC provides pedagogical mediation in the form of collocation learning tasks and corpus consultation: 1) the learning task draws the learner's attention to the gap in his/her knowledge of the target collocations; 2) corpora and concordancers are physical mediating tools whereby the learner generates and rearranges language in use for observation and analysis; and 3) the language data in corpora serves dual purposes: an object for observation and analysis, and a symbolic mediating means for restructuring the learner's word knowledge. In sum, through the mediation of various socially constructed auxiliary means (e.g., pedagogical tasks, corpora, concordancers, language in the format of concordances), DALC enables the learner to regulate collocation learning, namely appropriating the collocational patterns in real language (i.e., in the interpersonal plane) to reconstruct their mental lexicon (i.e., in the intrapersonal plane).

### **2.4.3 Combining the best of two worlds – Accounting for DALC with cognitive and sociocultural perspectives**

As noted earlier, DALC does not fit neatly into any single theoretical framework of SLA, it is most fully accounted for by a synthesis of cognitive approaches and sociocultural theory: SCT provides a sound basis for the pedagogical use of corpora and concordancers to *mediate* collocation learning; connectionism sees collocational knowledge as mentally represented as one of the connections of word knowledge; the connections within the mind are strengthened by repeated exposure to exemplars, which can be achieved by being exposed to a wealth of real language in corpora, albeit in a condensed, intensive manner (see 2.3.2.2); grounded in information processing approaches, the depth of processing hypothesis and the involvement load hypothesis provide theoretical underpinnings for inductive pedagogical approaches such as DALC; LLSs shed light on the induction processes in which the learner



cognitively manipulates corpus data in working memory as a prelude to *learning*.

I now turn briefly to consider the differences in L1A and SLA in order to demonstrate the value of corpus resources in L2 contexts. Sinclair (1991) conceptualizes language as organized on the *open-choice principle* and the *idiom principle* (see 2.1). Drawing on corpus evidence, he argues that (native) speakers adhere primarily to the idiom principle and only switch to the open-choice principle if certain constraints occur and prevent the former to function. While Sinclair's (ibid.) observation largely holds true insofar as native speaker data is concerned, Foster (2001) nonetheless found that nonnative speakers tended to construct a great proportion of L2 utterances from rules rather than from pre-constructed routines. This may be because that they do not have sufficient exemplars at their disposal to allow them to rely on the idiom principle. This sharp contrast in L1 and L2 production stems from different ways of learning L1 and L2. SLA differs from L1A<sup>5</sup> in many respects, most notably the *source* and the *volume* of language input, and L1 transfer (N. Ellis, 2003). L2 learners do not acquire the TL through naturalistic exposure, as L1 learners do: rather, they receive formal classroom instruction which is claimed by some to distort the patterns of exposure, of function, and of social interaction (N. Ellis and Laporte, 1997). Furthermore, adult L2 learners develop their TL knowledge on the existing L1 knowledge (MacWhinney, 1992), so L1 knowledge may have influences on L2 development to a certain extent, be it positive or negative. Acknowledging the differences between SLA and L1A, Randall (2007), on the other hand, does not see L2 learners as disadvantaged in terms of the learning environment and L1 influences. He argues that L2 learners have two relative strengths: 1) given a higher level of cognitive development, a L2 learner is more likely to examine language data and look for significant chunks in a more focused way; and 2) language in L2 classrooms is carefully structured in a way to highlight the significant chunks. Skehan (1998) suggests that L2 learners may avail themselves of higher levels of cognitive development to compensate for the lack of TL exposure in L2 contexts.

---

<sup>5</sup> Skehan (1998) conceptualizes L1A as a sequential process of *lexicalization*, *syntacticalization* and *relexicalization*. Exposed to naturalistic language input, L1 learners first learn language chunks of varying sizes (lexicalization), which are later subject to analysis and become the basis for extracting regularities (syntacticalization). The analysis may stop at this point or proceed to exploit the regularities to create other chunks which are available for future access as units of form-function composites (relexicalization).



As far as collocation learning is concerned, L2 contexts may not have as much naturalistic exposure as L1 contexts to allow learners to establish or strengthen collocational links incidentally, concordances can thus be an alternative, among others, to provide intensive, genuine TL exposure. Concordances are organized in such a way that language regularities become salient, so they can be usefully exploited to draw the learner's attention to the collocational patterns emerging from a profusion of language data. Also, DALC provides the learner with an opportunity to engage in 'cycles of analysis and synthesis' (Skehan, 1998) of corpus language as a means of inducing collocational patterns. As N. Ellis (2003) suggests, corpora of natural language provide the basis for a more systematic approach to the analysis of language, because they are the only reliable sources of frequency-based data. Given the limitations of L2 contexts and the affordances of corpus resources, DALC holds considerable promise for raising awareness of collocations and facilitating collocation learning in these contexts.

## **2.5 Empirical evidence of corpus-assisted language learning**

The previous section has considered the theoretical basis underpinning DALC, this section reviews the empirical literature on corpus applications in L2 learning contexts. A number of studies has discussed the ways in which corpus resources can be put into pedagogical use, and proposed a wide range of teaching and learning activities on various aspects of language, such as vocabulary (Stevens, 1991; Thurstun and Candlin, 1998); grammar (Qiao and Sussex, 1996; Whistle, 1999); translation (Bernardini, 2003); and EAP writing (Charles, 2007; Thompson and Tribble, 2001). While pointing to a promising direction for corpus-assisted language pedagogy, these studies have not empirically tested the pedagogical approaches and activities proposed therein. Cobb (1997, 1999) is among the first attempts to investigate the *effectiveness* of the learner's corpus consultation on vocabulary learning through experimental measures. Since then, much has been done to examine the efficacy of corpus consultation on various aspects of L2 learning: for example, vocabulary (Chan and Liou, 2005; Kaur and Hegelheimer, 2005; St. John, 2001; Sun and Wang, 2003); grammar (Belz and Vyatkina, 2008; Boulton, 2009; Chambers, 2005); translation (Bowker, 1999); and writing (Gaskell and Cobb, 2004; Kennedy and Miceli, 2001, 2010; Lee and Swales, 2006; O'Sullivan and Chambers, 2006; Todd, 2001; Yeh, Liou and Li, 2007; Yoon and

Hirvela, 2004). As the focus of this study is collocation learning assisted by corpus consultation, the following subsections review four relevant studies on the *product* (i.e., efficacy) of corpus consultation on vocabulary learning (see Section 2.5.1): Cobb (1999), Kaur and Hegelheimer (2005), Chan and Liou (2005) and Sun and Wang (2003); and three studies on the *processes* of corpus consultation (see Section 2.5.2): Kennedy and Miceli (2001, 2010) and Sun (2003).

### 2.5.1 The product of corpus-assisted vocabulary learning

Four studies, in particular, have examined the impact of learners' corpus consultation on their vocabulary learning: Cobb (1999), Kaur and Hegelheimer (2005), Chan and Liou (2005) and Sun and Wang (2003). The first two studies looked into the learning of *definitional knowledge* (i.e., form-meaning link) of words, while the other two studies focused specifically on *collocational knowledge*. More specifically, Chan and Liou (ibid.) explored the learning of one particular type of *lexical collocations*, *verb-noun collocations* (VN collocations), whereas Sun and Wang (2003) investigated the learning of *grammatical collocations*. The two studies on *definitional knowledge* of words are reviewed in Section 2.5.1.1, and the other two on *collocational knowledge* are taken up in Section 2.5.1.2.

#### 2.5.1.1 Effects on definitional knowledge

Cobb (1999) examined the effectiveness of corpus consultation on the learning of definitional knowledge of words. Participants were assigned to learn 200 words a week for 12 weeks. The target words were the 2,387 most common English words in Hindmarsh (1980). Control groups used a word list and dictionary, while experimental groups worked with a purpose-built corpus compiled from the learners' reading materials. A pretest, immediate post-test and delayed post-test along with weekly quizzes were administered to examine the learning outcomes. The results showed that in terms of *receptive knowledge* (as measured by multiple-choice questions), the two groups did not differ significantly in the immediate post-test. However, the learning effect of the experimental groups had persisted to the time of the delayed post-test, while that of the control groups had reduced. It seemed that the receptive knowledge of word definitions can be *gained* through exposure to word lists, dictionaries or concordances, whereas it can only be further *consolidated* through deeper processing

of language input induced by concordancing. On the other hand, the *controlled productive knowledge* of words was measured by gap-filling questions. The experimental groups significantly outperformed the control groups in this respect, which led the researcher to conclude that concordancing was conducive to the transfer of definitional knowledge to novel contexts. The findings attest to the assumption that successful learning of words can be promoted by meeting words in *varied situations* in addition to *varied contexts*, which is made possible by exploring corpus data. As Cobb (ibid.:348-349) notes:

[a] coherent text presents words in varied contexts but these tend to be limited to the few situations of principal concern to the writer, while a corpus is built from many texts and hence displays words in many more situations.

Cobb's (ibid.) findings demonstrate that observing a target word occurring in varied contexts and situations in the format of concordances may enable the learner to move a step further from being able to recognize the word, to recall it as the context demands.

While most experimental studies on the efficacy of corpus consultation on language learning looked into the learner's *receptive* and/or *controlled productive knowledge*, few ventured into *free production*, due to the relative difficulty of measuring this level of knowledge. Kaur and Hegelheimer (2005) was one of the few studies that examined the effects of concordancing on the learner's free production of academic words. The target academic words were selected from the 'Academic Word List' (Coxhead, 2000). 18 intermediate ESL learners were randomly assigned to an experimental group or a control group: the former was given access to a concordancer, 'Compleat Lexical Tutor' created by Tom Cobb, and an online dictionary, while the latter was only allowed to use the online dictionary. The results indicated that the experimental group made more attempts to use the target words and had more correct uses of them in the writing task. The percentages of correct word use differed significantly between the two groups: the experimental group scored 78%, while the control group achieved 67% ( $p < .05$ ). The findings suggest that receptive knowledge of word definitions gained from dictionary consultation is less successful when put to productive use. In contrast, a combination of dictionary and corpus consultation has more potential to transfer passive word knowledge (i.e., recognition) to active use (i.e., controlled or free production). Word knowledge gained in the form of dictionary



definitions is difficult to transfer to active use, possibly because this level of knowledge usually remains 'inert' (Cobb, 1999), while knowledge induced from multiple instances of actual use is relatively easier to put into production in novel contexts.

On Laufer's (1998) receptive-productive continuum of word knowledge (see 2.2.1), Cobb (1999) examined the effectiveness of corpus consultation on the *receptive* and *controlled productive knowledge* of word meanings, while Kaur and Hegelheimer (2005) investigated that on the *receptive* and *free productive knowledge*. Taken together, these two studies have demonstrated that corpus consultation enhances the definitional knowledge of words (i.e., form-meaning link) in all three levels: Cobb (1999) found that corpus consultation facilitated the acquisition and retention of *receptive knowledge*, and transferred it to *controlled production* in novel contexts; Kaur and Hegelheimer (2005) went further to illustrate that the transfer could be extended to *free production* as a result of concordancing.

### 2.5.1.2 Effects on collocational knowledge

The preceding section has discussed the efficacy of corpus consultation in facilitating the definitional knowledge of words. This section considers another aspect of word knowledge, collocational knowledge, and reviews two studies investigating the extent to which corpus consultation mediates the learning of collocations.

Chan and Liou (2005) investigated the effectiveness of corpus consultation on learning verb-noun collocations (VN collocations). 32 undergraduate EFL learners were given five sessions of web-based VN collocation instruction, of which three sessions were conducted inductively with the use of a Chinese/English bilingual corpus, while the remaining two sessions were carried out deductively without concordancing activities. A pretest, immediate post-test and delayed post-test were administered to examine the *learning effects* (i.e., changes that occurred between the pretest and immediate post-test), *retention effects* (i.e., changes that occurred between the immediate post-test and delayed post-test) and *residual effects* (i.e., changes that occurred between the pretest and delayed post-test). Also, a background questionnaire and an evaluation questionnaire were administered to gain insights into participants' perceptions of the web-based collocation instruction sessions. The findings indicated

that participants' performance on VN collocation tests improved significantly after receiving the web-based instruction, but regressed after an extended period of time (i.e., 2.5 months). Nonetheless, the performance on the delayed post-test was still significantly better than their entry level. More specifically, web-based collocation instruction with concordancing was more effective than that without concordancing in terms of extending participants' VN collocational knowledge and awareness-raising. Taking into account the participants' language proficiency, low-achievers benefited from the web-based collocation instruction to a greater extent than their high-achieving peers. With respect to their perceptions, while participants were in favour of explicit web-based collocation instruction, they did not have a positive attitude toward the use of the Chinese/English bilingual corpus. Interestingly, the participants did benefit from the use of this tool, notwithstanding their reservations. They were also reported to have encountered difficulties in using the bilingual corpus to generalize target collocations, yet acknowledged its value in collocation awareness-raising. Note that in Chan and Liou's (ibid.) study, collocational knowledge was measured by cloze questions, which in fact only elicited one mode of collocational knowledge, namely controlled productive knowledge. Therefore, the results here should be understood as: compared with the deductive instruction without concordancing, the inductive one with concordancing was more effective in extending the learner's controlled productive knowledge of VN collocations.

In the similar vein, Sun and Wang (2003) also looked into the efficacy of corpus consultation in developing collocational knowledge, except that the collocations under study here were the grammatical ones, as opposed to the lexical collocations investigated in Chan and Liou (ibid.). 81 EFL learners were randomly assigned to two groups, and respectively given an inductive pedagogical treatment with concordancing or a deductive one without concordancing in order to compare the differences in changes of collocational knowledge. The results demonstrated that the concordancing group made more progress than their non-concordancing counterparts in two respects: the overall knowledge of grammatical collocations and knowledge of *easy grammatical collocations*. However, the effects of concordancing did not seem to differ significantly from the deductive, non-concordancing approach with regard to *difficult grammatical collocations*. The researchers thus concluded that easy grammatical collocations were more amenable for the inductive approach with

concordancing, whereas the difficult ones could be facilitated by either an inductive or deductive approach. The knowledge of grammatical collocations was measured by error correction items which required two levels of knowledge: 1) identifying the error (i.e., receptive knowledge); and 2) correcting the error (i.e., controlled productive knowledge). As the researchers did not make a distinction between the two, there is no way of ascertaining whether the learner had difficulty in recognizing a miscollocation or correcting it when an item was marked wrong. Therefore, despite the claimed positive effects of concordancing on the overall knowledge of grammatical collocations, the extent to which concordancing facilitates the recognition and recall of grammatical collocations respectively remains unclear.

Taking together the findings of the four studies reviewed above, corpus consultation generally facilitates the development of various aspects of word knowledge (e.g., form-meaning links, grammatical and lexical collocations). However, aspects of word knowledge lie on a continuum from the receptive end to the free productive end, the aforementioned studies seem to provide only part of the picture, as they did not measure all levels of knowledge in that particular aspect. Many studies identified a discrepancy between the learner's receptive and productive proficiency of collocations (Bahns and Eldaw, 1993; Farghal and Obiedat, 1995), so all levels of knowledge should be taken into account when considering the learner's collocational knowledge. In this respect, Chan and Liou (2005) and Sun and Wang (2003) have gone only so far as to demonstrate the efficacy of corpus consultation in the learner's recognition and controlled production of collocations, so the extent to which the free productive collocational knowledge can be mediated by corpus consultation remains unresearched. In view of this, the present study examined the efficacy of corpus consultation in facilitating the learner's knowledge of lexical collocations from the *receptive, controlled productive* and *free productive* modes, so as to provide a comprehensive picture of the extent to which such endeavours restructure this particular aspect of word knowledge.

### **2.5.2 The processes of corpus-assisted language learning**

Though the efficacy of corpus consultation in facilitating various aspects of language learning has been extensively investigated, there is a paucity of empirical studies on



how learners actually fare in the course of corpus consultation, with the exception of Kennedy and Miceli (2001, 2010) and Sun (2003). These three studies are reviewed in this section, on account of their direct bearing on the present investigation of EFL learners' thinking processes during corpus consultation.

Kennedy and Miceli (2001) observed the corpus consultation behaviour of 10 intermediate FL learners of Italian and evaluated the effectiveness thereof. The evaluation carried out here was on the learner's corpus consultation skills, not on the learning outcomes that resulted from corpus consultation. Bearing in mind that much had been written on *what* can be done with corpora in language learning, the researchers explored *how* learners actually went about this endeavour.

The corpus used in Kennedy and Miceli's (ibid.) investigation was a small (57,000 words) purpose-built corpus 'Contemporary Written Italian Corpus' (CWIC), which was compiled from texts written by professional and non-professional writers. The participants were asked to undertake text-revising tasks with the aid of CWIC, while the researchers observed how they went about the tasks. In the light of their observation, the researchers generalized four linear steps of corpus consultation processes:

- 1) formulating the question;
- 2) devising a search strategy for a given question;
- 3) observing the data and selecting examples;
- 4) drawing conclusions.

The problems that hindered the completion of each step were identified, and suggestions for improvements were provided accordingly. Problems identified in the first step included failure to formulate a question that was explicit and searchable/manageable, failure to identify proper chunks and disregard of lexical considerations in favour of grammatical ones. When devising a search strategy in Step 2, learners tended to have difficulty in narrowing down the search scope to a manageable size and using more sophisticated functions to enhance the search. They were not flexible enough in situations where an alternative query word or form could have been used. Moving onto observing and selecting examples in Step 3, learners had difficulty in choosing appropriate examples that were illustrative of the target structure. To compound the problem, they tended to overlook something if it was not what they were looking for: that is, the search was often guided or even misled by the

presuppositions that learners had in mind. Finally, when it came to drawing conclusions, learners appeared to lack the confidence to make their own case based on the language evidence found, but allowed the number of search results to affect their decisions.

The summary of Kennedy and Miceli's (ibid.) findings above reveals that their study was *diagnostic* in nature. They identified novice corpus users' deficits in concordancing skills from the vantage point of expert users and thereby made suggestions for improving the concordancing skills. While their expert evaluation of the learners' concordancing skills did have some bearing on corpus-assisted language pedagogy, it may not have portrayed the learners' corpus consultation endeavours in the most rigorous way, considering that the diagnoses were made on the sole basis of *behaviouristic data* (i.e., the researchers' observation of the learner's corpus querying behaviour). In other words, their findings were primarily based on the researchers' observation and evaluation of the learners' corpus consultation behaviour, instead of the learners' self accounts of how they approached the tasks. Drawing on their observation of how learners interrogated the corpus, the researchers were able to identify the 'problems' encountered at each step over the course of the corpus consultation processes, and to provide corresponding 'tips' to increase the effectiveness of future corpus consultation endeavours. However, what can be observed from the learner's corpus consultation behaviour (e.g., queries entered onto a concordancer) is only part of the picture: the learner's failure to use certain preferred ways suggested by the researchers may have resulted from what underlay the observable behaviour, such as prior knowledge or learning style. Cognitive functioning or comprehension problems do not necessarily manifest themselves in query behaviour, and thus need to be elicited otherwise through *mentalistic measures* in addition to *behaviouristic* ones. Unless the full picture of this highly complex endeavour is revealed, pedagogical implications given may still remain on somewhat shaky ground. In view of this, the present investigation explored the learner's thinking processes during corpus consultation through both *mentalistic* and *behaviouristic* measures to provide a comprehensive picture of this endeavour (see 3.6). Moreover, the present inquiry was *exploratory* in nature, seeking to illuminate the ways in which the learner fared cognitively with corpus data, as opposed to the *diagnostic* investigation carried out by Kennedy and Miceli (ibid.).

Based on their previous study (i.e., Kennedy and Miceli, 2001), Kennedy and Miceli (2010) probed further into learners' corpus consultation endeavours. Instead of looking across learners to find a common pattern of corpus consultation behaviour (e.g., the four steps of corpus consultation presented above), as in the last study, the researchers carried out three in-depth case studies to explore how each individual fared with and perceived corpus consultation as a reference resource to assist creative writing. Also, the diagnostic evaluation of the learner's concordancing skills that was predominant in the previous study continued to play a part in this one as the grounds for improvements in their ongoing project.

Kennedy and Miceli (2010: 31-33) prescribed three functions by which a corpus served as a reference resource for creative writing: 1) *pattern-hunting*, which was defined as 'searching the corpus on words likely to be associated with the topic concerned, and then scanning the concordance lines for potentially useful patterns'; 2) *pattern-defining*, namely 'finding models when we do have a specific target pattern in mind for use at a particular point in a text. Usually it is a matter of knowing some of the component words and seeking a model for the exact structure required'; and 3) *finding an Italian equivalent for a given English pattern*. Instead of viewing the corpus consultation endeavour as a universal, linear process, as in the last study, the researchers shifted their attention to the idiosyncrasy in the ways in which each participant fared in such an endeavour. This paper provided a more detailed account of how each learner deployed the corpus functions above to solve their language problems and how they perceived their own undertaking in retrospection. Recognizing the highly idiosyncratic nature of corpus consultation endeavours, the researchers shifted from a micro-level to a macro-level to offer suggestions for developing corpus consultation skills. Whereas their previous study (Kennedy and Miceli, 2001) centered on the technical aspects of the corpus consultation behaviour, this one (Kennedy and Miceli, 2010) placed emphasis on more global issues of developing learners' *corpus consultation literacy*, as the researchers concluded, 'what matters is not mastery of a resource but of the function(s) it can be used for' (Kennedy and Miceli, 2010:40). As regards *corpus consultation literacy*, the researchers suggested that learners need to be sensitized to 'the limitations of reference resources, and one's own limitations in using them, with respect to specific functions, and specific problems they are used to address, rather than the resource itself' (ibid:41). The



present study also aims to uncover how learners go about corpus consultation through descriptive accounts. While Kennedy and Miceli's (ibid.) findings showed how their learners deployed the prescribed corpus functions to solve language problems, the present study went a step further to demonstrate how the learner's mental capacities came into play as an integral part of the consultation process and how they synchronized the innate capacities with physical and symbolic mediating tools (see results in Chapter 5).

Sun (2003) looked into the corpus consultation behaviour of three intermediate EFL learners. With the assistance of corpora and concordancers, the participants were asked to proofread a list of eight questions with various types of grammatical errors. They were also required to think aloud simultaneously as they undertook the proofreading task. The results showed that four factors influenced the learning outcomes of corpus consultation: *prior knowledge*, *cognitive skills*, *teacher intervention* and *concordancer skills*. Sun (ibid.) further subcategorized *cognitive skills* into four stages - comparing, grouping, differentiating and inferring. These stages were not distinct, and they may overlap depending on how the learner approached the task. Both Sun (ibid.) and the present study adopted a mentalistic, introspection measure, namely think-aloud, to generate underlying cognitive data that would not have been obtained otherwise, but this study further complemented mentalistic data (i.e., think-aloud verbal protocols) with behaviouristic data (i.e., corpus queries) so as to reconstruct the learner's thinking processes as closely as possible. The cognitive skills identified in Sun (ibid.), along with O'Malley and Chamot's (1990) taxonomy of LLSs (see 2.4.1.2), served as the preliminary framework for the present inquiry into EFL learners' thinking processes during corpus consultation.

## 2.6 Summary

This chapter has characterized collocation from the theoretical and corpus linguistic perspectives, highlighting its ubiquitous and usage-based nature. The classificatory discussion has set out the full scope of formulaic language so as to delineate the focus of this study, namely *lexical collocations*. Moving from linguistic to pedagogical issues, the importance of developing collocational knowledge in L2 learning can be

argued from multiple perspectives: collocational knowledge forms the basis of language knowledge; collocational knowledge is a key component of lexical knowledge; and collocational knowledge is attributable for the accuracy, fluency and appropriateness of language production. The ubiquitous and usage-based nature of collocations is nowhere more evident than in corpora. Corpus resources hold considerable promise for facilitating collocation learning in L2 contexts where naturalistic exposure to the TL is lacking. To understand the impact of the pedagogical use of corpora on collocation learning, a corpus-assisted approach, DALC, has been critically discussed. As DALC is a relatively new pedagogical approach, it does not fit neatly into any single SLA theoretical framework. Having reviewed the SLA literature, I argue that the theoretical foundation for DALC is one that synthesizes cognitive approaches and SCT to underpin such a complex pedagogical action. SCT provides a sound basis for the pedagogical use of corpora as a mediating means. The depth of processing hypothesis and the involvement load hypothesis underpin inductive pedagogical approaches such as DALC. Connectionism sheds light on the mental representation of collocational knowledge. LLSs, grounded in information processing approaches, account for how collocation learning occurs as a result of cognitively manipulating corpus data and appropriating it into the mental lexicon. I then shifted the focus from theoretical perspectives to empirical ones, reviewing the relevant studies on the *product* and *processes* of corpus-assisted language learning. Having reviewed the theoretical and empirical literature, I will discuss the methodological issues as to how this inquiry into DALC has been conducted in the next chapter.

## Chapter 3 Methodology

In order to address the focus of this research, namely the *learning product*, *learning processes* and *learner perceptions* of DALC, a principled investigation was carried out. This chapter details the research methodology employed in the investigation. It begins with the philosophical stance taken to guide this research inquiry (3.1). The aims and overall design of the research are set out in Section 3.2. The sampling and characteristics of the participants are taken up in Section 3.3. Section 3.4 outlines the pedagogical mediation, DALC. The data collection methods and instruments, and data analysis methods used to address the three research questions are discussed in Sections 3.5-3.7, followed by validity and reliability issues (3.8) and ethical considerations (3.9). The chapter concludes with a summary of the research methodology (3.10).

### 3.1 Philosophical stance

A philosophical stance is 'the basic belief system or worldview that guides the investigator, not only in choices of method but in ontologically and epistemologically fundamental ways' (Guba and Lincoln, 1994:105). The present study was guided by an eclectic philosophical stance: ontologically and epistemologically committed to *constructivism*, while methodologically informed by *positivism*. Ontologically, what underlies my view on *language* and *learning* is the *constructivist* belief in reality as the intersubjective co-construction of an individual and society, rather than an objective entity 'out there', independent of the knower. Epistemologically, I accept the *constructivist* posture that knowledge, and language in particular, is a socially constructed artefact. On the other hand, whereas *positivism* may not adequately account for the immense complexity of the phenomena in the social sciences, its established research conventions and emphasis on the rigour of inquiry nonetheless have important bearing on educational research methodology, including that of the present study.



### 3.1.1 Ontological and epistemological commitment to constructivism

Central to constructivist ontology is its opposition to *objective reality*, an ontological view stemming from *realism*. Constructivism relinquishes the realist belief in the existence of a world 'out there', independent of an individual's awareness or understanding. Instead, it holds that reality is the result of human construction influenced by and embedded in social contexts. There is no observed phenomenon without an observer, or as Piaget (1937 cited in Glasersfeld 1989:136) puts it, 'intelligence organizes the world by organizing itself'. Lincoln (1990:77) elucidates the ontological commitment of constructivism by contrasting it with that of positivism:

[t]he ontological axiom [of constructivist inquiry] states that reality is a social, and, therefore, multiple, construction, that there is no tangible, fragmentable reality on to which science can converge, that reality exists rather as a set of holistic and meaning-bounded constructions that are both intra- and interpersonally conflictual and dialectic in nature, that, whereas the positivistic construction of reality is realist in orientation, the constructivist is relativists, that whereas the aim of positivistic science is to expose and articulate immutable natural laws (for both the social and the natural world), usually expressed as generalizations, and usually in the form of cause-and-effect relationships, the aim of constructivist science is to create idiographic knowledge usually expressed in the form of pattern theories, or webs of mutual and plausible influence expressed as working hypotheses, or temporary, time- and place-bound knowledge.

My ontological commitment to constructivism, particularly *social constructivism*, is nowhere more evident than in my view of *language* as a socially constructed artefact, which does not exist *a priori* independently of its speakers, but is a product of human collaborative endeavours in a particular speech community. Since language is the artefact co-constructed by its speakers, it is by no means *objective* but *intersubjective*. In view of this, the criteria for judging language knowledge within the constructivist paradigm are *intersubjectivity* and *viability* rather than *objectivity* and *reality/truth* embraced by positivism. In other words, constructivists do not judge the validity of one's knowledge by the extent to which it mirrors reality, but the extent to which it accords with the construction on which most people of a social group agree (i.e., *intersubjectivity* embraced by social constructivism) or the extent to which it is a viable action relative to other alternatives (i.e., *viability* proposed by radical constructivism) (Duffy and Cunningham, 1996). Glasersfeld (1989:134) aptly describes the notion of *viability*: '[i]nstead of presupposing knowledge is a representation of what exists, knowledge is a mapping, in the light of human experience, of what is feasible'. Such a *relativist* thinking is highly pertinent to my

view on the linguistic phenomenon of collocation: collocability is not judged against a right-or-wrong dichotomy but a cline of probability of being accepted/used among speakers in a particular speech community. Also, as language is co-constructed and used by people of a social group, collocability cannot be judged in a vacuum: the context in which a collocation is embedded needs to be taken into account (see 3.5.3.4 for judging the acceptability of learner collocations).

Related to *anti-realist* ontology is constructivist epistemology that stresses the process of co-constructing knowledge. Constructivism is 'a theory of knowledge acquisition that sees learners constructing their own knowledge and meanings on the basis of personal experiences' (Firth and Wagner, 2007:806). Along the same lines, Ruschoff and Ritter (2001) argue that constructivist learning encourages learners to actively cognize with the assistance of resources, rather than passively receive formal instruction from the teacher. In contrast, from a realist viewpoint, the goal of learning is to gain knowledge about the world of *objective reality*: what the learner receives is the same as what is transmitted from the teacher. Learning, according to a positivist/realist viewpoint, consists of replicating *objective reality* in the learner's mind (Jonassen, 1991). However, knowledge acquisition within the constructivist paradigm is not gaining knowledge of *objective reality*, but engaging in the co-construction of an intersubjective artefact embedded in a particular socio-cultural context. As Glasersfeld (1987:16) states, '[k]nowledge is not a transferable commodity and communication not a conveyance'. Within the constructivist framework, learning is characterized by the learner actively engaging in making sense of his/her own experiential world and constructing meaningful representations. It is therefore the *process* rather than the *product* of learning that plays the pivotal role in knowledge construction. Confrey (1990:109) goes on to claim that constructivism not only underscores the constructive process, but also emphasizes the active role we take to be 'at least partially able to be aware of those constructions and then to modify them through our conscious reflection on that constructive process'. The present study was guided by constructivist epistemology: language learning is a process in which the learner actively and collaboratively constructs knowledge by interacting with the target language as well as speakers of that language.

### 3.1.2 Methodology informed by positivism

Though the present study was ontologically and epistemologically guided by *constructivism*, it nonetheless employed a *positivist* methodology. A different view of what constitutes reality and knowledge does not necessarily lead to a rejection of *rigour* in data collection and analysis. Recognizing the fundamental differences in the ontological and epistemological commitment between the two paradigms, I adopted an eclectic approach to investigating corpus-assisted collocation learning, looking into multiple dimensions thereof (i.e., the *product*, *processes*, and *learner perceptions*) using a positivist methodology.

Positivist methodology is characterized by empirical verification of knowledge through objective manipulation and control over variables. Such inquiries are concerned with *causal inferences*, which can only be verified by isolating extraneous variables from the investigation. Any variables that may confound the causal relationship under study should be carefully controlled, in particular the inquirer's value involvement. *Objectivity* needs to be guarded at all costs by avoiding subjective judgment. However, as far as education is concerned, a total exclusion of value seems unlikely, because education itself is an enterprise undertaken by and for human beings. As Carr (1995:97) argues, 'educational research always involves a positive commitment to educational values, the pursuit of objectivity as understood in the natural sciences is undesirable'. Along the same lines, Kirk and Miller (1986:20) provide a more adequate conceptualization of *objectivity* in the social sciences: 'objectivity is the simultaneous realization of as much reliability and validity as possible'. Indeed, from a constructivist viewpoint, the language learning phenomenon this study set out to investigate (i.e., collocation) was in itself *constructed reality*, as opposed to *objective reality*. Even the ways in which the research was designed and the findings were interpreted were inevitably constructed, so a wholesale abandonment of *value* was unlikely, if not impossible. *Objectivity* was nonetheless guarded carefully throughout this inquiry process, bearing in mind that *objectivity* here was understood in Kirk and Miller's (ibid.) sense, rather than that of the natural sciences.

It is noteworthy that a common misconception about positivism is that it is associated



exclusively with quantitative research. Crotty (1998) argues that qualitative research can be undertaken within the positivist framework and non-positivists also employ quantitative methods. A philosophical stance informs, but by no means limits the choice of research methods. Notwithstanding the fact that positivism has been discredited in some educational research areas, it is unlikely to totally avoid positivistic research conventions as far as the ultimate aim of improving educational practice is concerned. As Schrag (1992:7) states:

[i]nsofar as any research program aspires to enhance educational practice, it must ultimately issue in some policy, way of thinking, conceptual framework, design, strategy, or practice for intervention in the lives of children. At that point, it is incumbent to ask whether the intervention is an improvement on current practice. If the argument is to be persuasive, it must show the superiority of the innovation. To demonstrate that superiority, it will have to provide evidence that compared with current practice the innovation yields more educational value. Where can such evidence come from? It can come from philosophical considerations that support or undermine the innovation regardless of its consequences. Or it can come from data derived from experiments that utilize the educational trial. I see no other alternative.

The present study aims to investigate three key dimensions of DALC: the *learning product*, *learning processes* and *learner perceptions* (see 1.3 for research questions). The first key aim is concerned with the causal relations between DALC and the changes that occur in the learner's collocational knowledge as a result of receiving DALC. To explore the nature of such causal relations, a *pretest-post-test non-equivalent group design*<sup>6</sup> was employed so that the effects of DALC intervention could be made clear through within-group and between-group comparisons (see the research design in 3.2). Extraneous variables such as participants' prior collocational knowledge were controlled to verify the causal inference (also see 3.4 and 3.9 for discussion on the dilemma between positivistic control in naturalistic research settings and ethical considerations). In addition to the learning outcomes, the study also looked into the processes leading to such outcomes: mentalistic data as well as behaviouristic data were gathered to reconstruct the thinking processes as the learner undertook the DALC task. I was fully aware that the researcher's reconstruction of the learner's thinking processes inevitably entailed a certain degree of subjectivity, be it the ways in which the processes were categorized or the results were interpreted and reported. However, it was with this awareness in mind that I proceeded with caution, analyzing data and reporting findings as truthfully as possible (see 3.6.3 for the

---

<sup>6</sup> This design consists of an experimental group and a control group, both of which take a pretest and a post-test. But the two groups are not equated by randomization, hence the term 'non-equivalent' (Cohen, Manion and Morrison, 2007).

measures taken to guard against threats to the reliability of this dataset). Also, this study examined learner perceptions of DALC with a questionnaire, a positivistic approach which operationalized psychological constructs into numerical values. Again, my interpretation of questionnaire responses was constructed on my understanding of the results, and thus was not completely value-free. Acknowledging the inevitable value involvement in education research, I endeavoured to maximize the rigour of this research by realizing as much reliability and validity as possible (as will be critically discussed in 3.8).

## 3.2 Research aims and design

As Mouly (1978:12) aptly puts it,

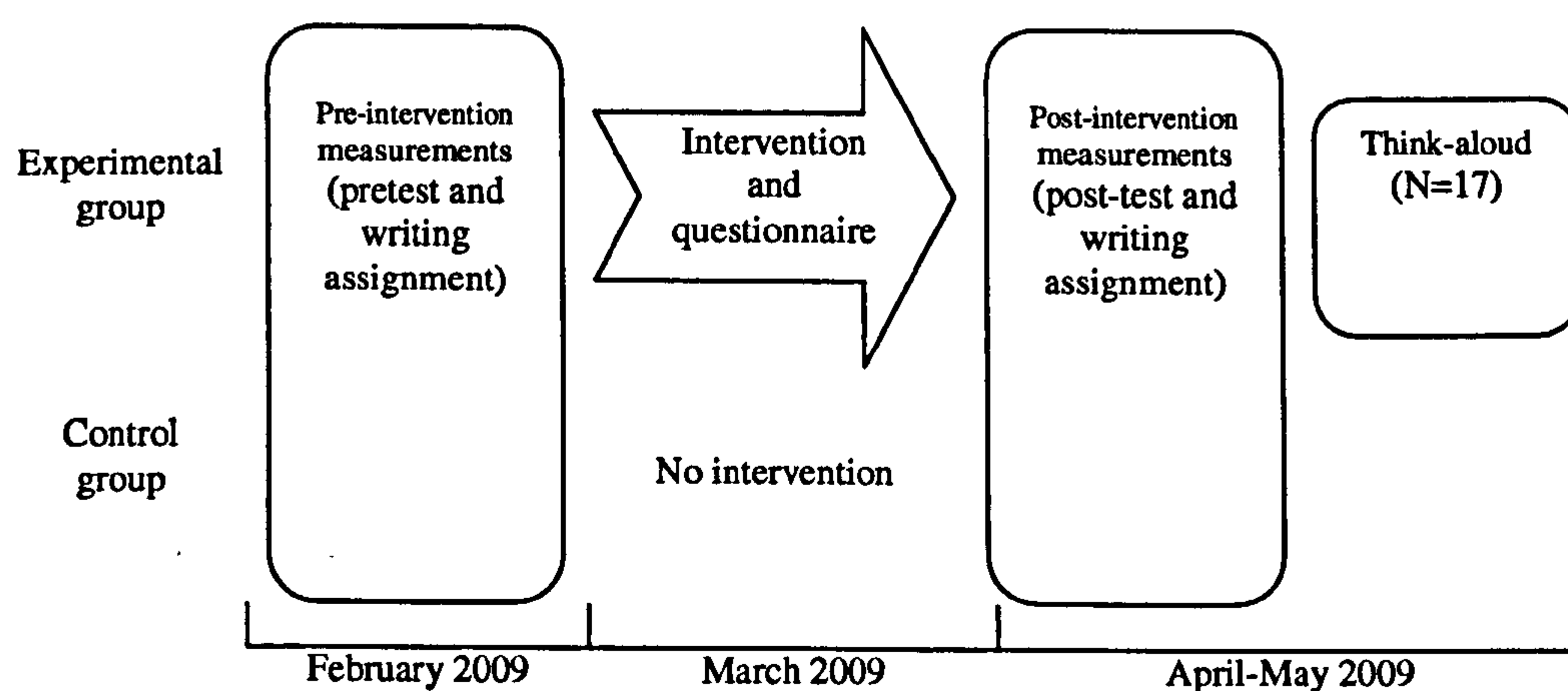
[r]esearch is best conceived as the process of arriving at dependable solutions to problems through the planned and systematic collection, analysis, and interpretation of data. It is a most important tool for advancing knowledge, for promoting progress, and for enabling man to relate more effectively to his environment, to accomplish his purpose, and to resolve his conflicts.

My observation of the discrepancy between the principle and the practice of collocation learning in EFL contexts motivated this research quest for way(s) to develop EFL learners' collocational awareness and knowledge. In addition, corpus resources have been heralded for their potential to reveal linguistic regularities, of which collocation is one kind. In view of the potential corpus resources have for developing collocational knowledge, this study set out to explore how collocation learning can be mediated by learners' self-access corpus consultation. The current research looked into three key dimensions of DALC: the *product*, *process* and *learner perceptions*. The key aims were formulated into three corresponding research questions:

- RQ1: Does a data-driven approach to learning collocations facilitate EFL learners' development of collocational knowledge? If so, how does it facilitate such development?
- RQ2. What is the nature of the thinking processes EFL learners engage in during the data-driven approach to learning collocations?
- RQ3. How do EFL learners perceive the data-driven approach to learning collocations?

A multidimensional investigation into the data-driven approach to learning collocations promises to provide a comprehensive picture of how collocation learning

takes place through the mediation of corpus exploration. To address RQ1, measurements on the participants' *collocational knowledge* were taken before and after receiving DALC intervention to examine the *learning outcomes*. *Collocational knowledge* was measured at three levels: *receptive*, *controlled productive* and *free productive knowledge*. The instruments were collocation tests, worksheets and writing assignments (see 3.5 for the methods for RQ1). In response to RQ2, a small group of participants (N=17) were sampled to do *concurrent think-aloud interviews* in order to explore the *thinking processes* as they approached the DALC task. *Mentalistic data* (i.e., think-aloud verbal protocols) as well as *behaviouristic data* (i.e., corpus queries) were collected to reconstruct the respondents' thinking processes (see 3.6 for the methods for RQ2). To address RQ3, an evaluation questionnaire was administered to elicit *learner perceptions* of DALC in relation to their vocabulary learning experience and collocational awareness (see 3.7 for the methods for RQ3). The aforementioned methods of data collection and analysis will be discussed at length in the following sections (3.5-3.7). The data collection procedure in the main study is visually represented in Figure 3-1:



**Figure 3-1** The data collection procedure

As illustrated in the figure above, this study adopted a *pretest-post-test non-equivalent group design* in which the experimental group received DALC intervention, while the control group did not, so as to provide a baseline for comparisons. Measurements of collocational knowledge were taken from both groups before and after the intervention or non-intervention period, for within-group and between-group comparisons. The subsequent sections detail the methods used to gather and analyze



data on the *learning product*, *learning processes* and *learner perceptions* of DALC.

### 3.3 Participants

The participants of this study were 186 undergraduate EFL learners majoring in English at a university in Taiwan. This was *convenience sampling*: that is, the sample was chosen from those to whom the researcher had easy access, because randomization as in a true experimental design was not feasible (Cohen, Manion and Morrison, 2007). This sample came from four intact classes to which they had been assigned prior to the study. Two classes were then randomly assigned to the experimental group (N=109) and the other two constituted the control group (N=77). Such a quasi-experimental design using intact classes is favourable in many educational research settings, because it causes less interruption to the existing school system (Porte, 2002). Compared with true experimental design, quasi-experimental design is more realistic and amenable to generalization to a wider range of educational contexts. As some data in the study was quantitative in nature (e.g., test scores or questionnaire responses), sample size played an important part in determining the generalizability of the results yielded. A group size of approximately 100 experimental participants was large enough to allow statistical computation, yet manageable, considering the computer facilities required for DALC intervention.

This cohort was highly homogeneous in that all participants were Taiwanese and native speakers of Mandarin Chinese. English is a foreign language in Taiwan, so most of the input is received from formal tuition at school. English is a compulsory subject within the curriculum of secondary education in Taiwan, which comprises three years of junior high school and three years of senior high school or vocational high school. Therefore, participants were assumed to have received a minimum of six years of formal tuition of English by the time they enrolled at the university. In addition, one of the graduation thresholds of this department is that the students must reach either of the following standards by the time they finish the four-year curriculum:

TOEFL CBT: 197 (PBT 527)  
TOEIC: 750  
IELTS: 5.5

A band score of 5.5 on IELTS represents the language proficiency between a modest user and a competent user. Since the students of this department are expected to reach such language proficiency after four years of training, it follows that they may be well below such a proficiency level in their first year. As all the participants were first-year students, they were assumed to have lower-intermediate levels of English proficiency.

All first-year undergraduates at this university are required to take the course 'General English' as part of the first-year curriculum, English majors are no exception. DALC intervention was implemented as an addition to this course, because it was a required course taken by both groups of participants. The participants were instructed by the same teacher in this course. The following section will give a detailed account of DALC intervention given to the experimental participants. The coursebook selected was *Visions B: Language, literature, content* (McCloskey and Stack, 2004), which was designed for ESL/EFL learners of lower-intermediate to intermediate proficiency, namely those who score as follows on either of the three major standardized tests.

TOEFL iBT: 61-80 (CBT: 173-213)  
TOEIC: 650-750  
IELTS: 4.6-6.0

Based on their average length of formal study of English, the language requirement of the curriculum and the level of the coursebook chosen, the participants were assumed to be at lower-intermediate levels of English proficiency.

### **3.4 DALC intervention**

DALC intervention, consisting of five weekly sessions, was given to the experimental group from 5<sup>th</sup> March 2009 to 2<sup>nd</sup> April 2009. Meanwhile, the control group did not receive DALC intervention, but the teacher would introduce the target collocations as part of her vocabulary pedagogy. As mentioned in the previous section, there was a coursebook used in the course 'General English', where the treatment for both groups was implemented. As the coursebook is mainly designed for reading and contains texts of various genres, the pedagogy adopted by the teacher was predominantly grammar translation: the teacher would have the students read aloud a paragraph in the designated text and then translate it sentence by sentence, adding explanations of vocabulary and grammar rules. By explaining vocabulary I mean providing students

with the definitions, grammatical functions, synonyms/antonyms and collocations. In particular, to ensure that the control participants were also exposed to the target collocations as their experimental counterparts, the target collocations would be introduced to the control participants by writing them down on the blackboard. In contrast, as will be elaborated below, the experimental participants were not taught the target collocations directly by the teacher, instead, they were given corpus resources to induce target collocations by themselves. In the case of the control treatment, while the class time (50 minutes) was approximately divided into 20 minutes of vocabulary teaching and 30 minutes of grammar teaching, collocation teaching usually took a small part of the vocabulary teaching time (approximately 5 minutes). As for the experimental treatment, although they were given 30 minutes to work exclusively with corpus resources to induce target collocations, they were not exposed to the target collocations for the entire 30 minutes: they had to spend most of the time searching and generalizing target collocations from a multitude of corpus data. In this case, it is unlikely to control the actual amount of time an experimental participant was exposed to target collocations, because participants would spend different amounts of time inducing even the same target collocation, depending on the corpus resources consulted, corpus consultation skills or prior knowledge. Unlike the control treatment which was more direct but took a shorter period of time, the experimental one, though longer in time, was predominantly learner-centred and less structured.

Admittedly, the distinction between the two groups was not as tightly controlled as that in a true experimental study, where all the potential confounding variables were either excluded or controlled. In other words, I had no control for all the input the learner might have had from the curriculum or outside the language classroom. On the other hand, I had no intention of imposing such stringent experimental control in this naturalistic research setting, because the research aim was to understand how DALC differed from the usual learning conditions the participants were in, so apart from DALC intervention given to the experimental group, the learning conditions of all the participants remained as naturalistic as possible (also see 3.9 for ethical considerations).

Prior to the intervention, a training session that entailed instructions and a hands-on trial was given to the experimental group, in order to familiarize the participants with



the skills required to work with web-based concordancers (see Appendix 4 for the handout distributed in the training session). Each intervention session took approximately 30 minutes after the 'General English' class. In the sessions, the participants were given collocation learning worksheets (see 3.5.2.3 for designing the worksheets, and Appendix 9 for the worksheets) along with computers and Internet access, and were instructed to use six designated web-based concordancers as an aid to search for the target collocations on the worksheets. For example, the following question asked the learner to identify one miscollocation from four options:

highly educated

highly exhausted

highly profitable

highly unusual

The learner may interrogate corpora for the frequent adjective collocates of the adverb *highly*:

the Deacons/Hendricks. Her work is very bold records of a specific site, and ends with scary, out there, completely compelling and better but above all designer of his own work. and a convulsive, irresistible chorus. It's Hee-la), with its steep rocky slopes, dense in the American congress which has been s has himself conducted what's been seen as a are some jobs in which people have access to in the first place. [o] They were of a nature of the [f] Chronicon [f] It is in hunting kit with shiny brown boots and that, with the most sincere and in many ways for the subject. By shifting attention to the Snake is born under the sign of wisdom. He is [p] Polarization can make sense in the for administering large quantities of designer Stuart Devlin whose work has become Surrealist era. Over the years she evolved a Mr Humphries. "This is our forte, making these cent. [p] The Treasury's growth forecast is BOOK 19 [p] STORY LINE [p] KRYSALES, a round the corner with a gun. Over here it's her bankers, Coutts. [p] Status [p] She has a appreciation have combined to generate some further gratification. Smearing faces proved to have the input and assistance of several and elsewhere. The common denominator in these more precise. Carter assumed that there were n twins reared together have ids that are The quality of care appears to be a Her high school in the nation's capital was yield, high-risk bonds gave it room to offer 1.7 tons of plutonium. Japan plans to use the eat. This scoop, invented back in 1935, has a not only bad for our politics it also is also highly coloured. She always asks me for a ruler. highly Romantic works from the mid-19th century. It highly original. Listen and be changed. [p] DONNA highly individual, it could sometimes be described highly unlikely, however, that radio programmers ( highly flammable vegetation, and unpredictable highly critical of both China's crackdown on pro highly effective cross-examination of witnesses. highly sensitive information. What we would say is highly personal nature and the newspaper had highly unlikely that he would have mangled it so highly polished gaiters, a [f] maradadi [f] maradadi highly commendable motives, our Director is highly fluid and complex nature of missionary highly intelligent and his mind is forever active. highly competitive world of overseas consultancy highly toxic substances, which demands considerable highly collectible. A stunning candelabra by Devlin highly effective technique for conveying her highly complex textiles in the same manner as when highly dependent on a bounce in consumer spending highly sensitive Mato file, has gone missing from highly likely." [p] Wandering around New York highly extravagant lifestyle which she justifies by highly attractive rates of return. And keep in mind highly effective as a method of fleeing the highly talented people without whose guidance and highly disparate societies seems to be not only highly unlikely steps involved in the evolution of highly similar, much more similar than what occurs highly significant element. Good-quality care highly intellectual in comparison to the one in highly competitive rates. But now, the multibillion highly toxic, radioactive fuel in a serious of highly polished, one-piece aluminum handle that highly inefficient. And he has talked about how

Figure 3-2 Concordances of 'highly' from the Collins WordbanksOnline English Corpus

The designated web-based concordancers were *Collins WordbanksOnline English Corpus*, *IWill Collocation Explorer*, *Lexical Tutor*, *NTNU Web Concordancer*, *TOTALrecall* and *VLC Web Concordancer* (see Appendix 5 for further information about the concordancers). All the concordancers were freely available online, and thus provided a good starting point for learners who wished to exploit such tools to assist language learning. By observing the output generated by concordancers, the participant was able to form and test linguistic hypotheses, and eventually reached generalizations about the target collocational patterns. It was hypothesized that through such intensive exposure to real language data and induction of frequent collocations, the learner would be sensitized to the typical patterns, contextual preferences and frequency distributions of the target collocations.

### 3.5 Collecting and analyzing data on learning product

Measurements were taken to understand the effects of DALC intervention noted in the previous section. This section details the data collection and analysis methods employed to address the first research question:

RQ1: Does a data-driven approach to learning collocations facilitate EFL learners' development of collocational knowledge? If so, how does it facilitate such development?

The effects of DALC were examined from the *changes* that occurred in the learner's collocational knowledge before and after the intervention or non-intervention period (i.e., within-group comparisons) and the *differences* in collocational knowledge between the experimental group and their control counterparts (i.e., between-group comparisons).

#### 3.5.1 Collecting data on collocational knowledge

As noted in Section 2.2.1, lexical knowledge falls on a receptive-productive continuum, rather than exists as an all-or-nothing dichotomy. In other words, lexical knowledge should be regarded as *continuous* instead of *dichotomous*. Collocational knowledge, being one aspect of lexical knowledge, also operates along a continuum. Drawing on Laufer's (1998) levels of vocabulary knowledge, collocational knowledge is conceptualized as a collective construct, composed of *receptive knowledge* (i.e., recognition), *controlled productive knowledge* (i.e., cued recall) and *free productive knowledge* (i.e., free recall). Therefore, it is only appropriate to measure learners' collocational knowledge at all three levels so that a comprehensive picture may be obtained.

*Receptive collocational knowledge* in the present study was measured with multiple-choice questions (henceforth MC questions), *controlled productive knowledge* was operationalized as the performance on gap-filling questions, and *free productive knowledge* was examined through learners' actual use of collocations in their writing assignments. Most previous studies only measured one or two levels of collocational knowledge, in particular receptive and/or controlled productive knowledge, for ease of item construction and validation. Few have ventured to look

into the learners' free productive knowledge of collocations. In order to understand the full extent of the impact of DALC on collocation learning, this study gathered data from three levels of collocational knowledge to provide a more complete picture of the efficacy of DALC.

To examine the changes in the learners' collocational knowledge brought about by DALC, measurements were taken at two points in time: prior to and subsequent to the intervention or non-intervention period (see Figure 3-1 for the visual representation of the data collection procedure). Pre-intervention measurements consisted of a pretest of collocations with MC questions and gap-filling questions (see 3.5.2.1) and writing assignments written before the (non)intervention period (see 3.5.2.2). Post-intervention measurements entailed a post-test (see 3.5.2.1) and the writing assignments after the (non)intervention period (see 3.5.2.2). Note that the measurements were taken from both groups, but only the experimental group received DALC intervention. Also, the worksheets distributed to the experimental participants in the intervention sessions were collected to provide formative information on corpus-assisted collocation learning.

### **3.5.2 Instruments eliciting data on collocational knowledge**

This section outlines the instruments used to elicit data on collocational knowledge, including *tests* (receptive and controlled productive knowledge), *writing assignments* (free productive knowledge) and *worksheets* used in the DALC intervention sessions.

#### **3.5.2.1 Collocation tests**

A pretest and a post-test on target lexical collocations were administered to measure the changes in the learners' receptive and controlled productive collocational knowledge brought about by the intervention. In order to control the level of test difficulty, the items in the pretest and post-test were approximately the same, with slight modifications on MC questions and a different sequencing of items. The same set of items was also used in the collocation learning worksheets (see 3.5.2.3) distributed in the DALC sessions. This section gives a procedural account of how the collocation tests were developed.



## **1) Selecting target collocations**

The base words of the target collocations in the tests were drawn from the wordlist in the prescribed coursebook *Visions B: Language, literature, content* (McCloskey and Stack, 2004). Instead of distracting participants with collocations of a new set of words, the present study used the prescribed wordlist in their coursebook with the aim of extending learners' knowledge of these words by introducing their frequent collocations. Since the prescribed words had been studied prior to the time of this study, the participants were assumed to have had knowledge of these words, at least at the level of word meanings (i.e., form-meaning links). It was thus hypothesized that DALC intervention on the frequent collocations of these previously studied words would enhance the learner's word knowledge by establishing collocational links on the basis of definitional knowledge (see 2.4.1.1 for connectionist account of learning as establishing connections).

The instructor of the course and I selected 50 base words, on the basis of whether the word had a range of frequent collocations that might be useful for the participants to learn. As the present study looked into the learning of lexical collocations, leaving out the grammatical ones, only the lexical collocations of the 50 base words were identified with reference to *The BBI Dictionary of English Word Combinations* and *Oxford Collocations Dictionary for Students of English*.

## **2) Constructing and piloting the preliminary test**

To develop collocation test items, I searched online for authentic texts in which the candidate collocations noted above were embedded. Texts that had the candidate collocations and were comprehensible to the participants were excerpted, and developed into test items. The excerpts were adapted as necessary to provide more contextual clues or to fine-tune the linguistic difficulty. The (adapted) excerpts were examined by the course instructor to ensure that the level of difficulty was suitable for the participants. Based on the (adapted) excerpts, fifty items for the preliminary test were constructed in the format of MC questions and gap-filling questions (see Appendix 6 for the preliminary test).

The preliminary test was piloted to ensure its reliability, validity and practicability. It

was administered to a closely matched sample of 89 participants on 22<sup>nd</sup> December 2008. The participants in the pilot study were comparable to those in the main study in that they were also undergraduate EFL learners majoring in English who had a minimum of six years of formal study of English. The reliability of the preliminary test was .711 (Cronbach's alpha), which, according to Field (2005), indicated a good level of consistency of the measure as far as ability tests were concerned. The preliminary test was then revised based on the results of the pilot study. Most of the items in the preliminary test were retained except for four gap-filling questions. One item was excluded because no participant could answer it. The other three items were removed because the target verb collocates were all *make* (i.e., *make an interpretation*, *make a diagnosis* and *make an observation*). The responses to such items could be confounded by guesswork or a strategic use of de-lexicalized words, rather than the learner's actual collocational knowledge. The items above were replaced by four new items testing the following lexical collocations: *a sense of achievement*, *deep devotion*, *vivid imagination* and *keen observation*.

### 3) The pretest and post-test in the main study

As noted earlier, the pretest and post-test were similar, with slight differences in MC questions and sequencing. The pretest had 23 MC items and 27 gap-filling items, and the post-test had 22 MC items and 28 gap-filling items (see Appendices 7 and 8 for the pretest and post-test). One MC item was converted into a gap-filling one for ease of corpus search.

Each MC question had three options. The rationale for using MC questions was twofold: 1) to measure the learner's receptive collocational knowledge; and 2) to raise the learner's awareness that collocations were largely subject to actual usage rather than semantics (see the arbitrariness in the choice of collocates in 2.2.2.2), so all the options might be semantically similar but only some were typically used. To reduce *practice effects*, the MC items were slightly modified in the post-test: MC questions in the pretest consisted of two correct options and one incorrect option, while those in the post-test had one correct option and two incorrect ones. For example:

Pretest: The professor \_\_\_\_\_ a speech about current economic situations.  
(Choose one INCORRECT option)  
(A) delivered (B) gave (C) sent

Post-test: The professor \_\_\_\_\_ a speech about current economic situations.  
(Choose one CORRECT option)  
(A) delivered (B) sent (C) talked

A correct option here refers to the frequent and typical collocate of a base word. Whereas the incorrect options are grammatically well-formed and semantically similar to the correct option, they do not qualify as acceptable collocations in the present study because they are not significantly typical of how the base word expects its collocates<sup>7</sup>. Lewis (2000) argues that showing false/unacceptable collocations, in contrast to appropriate/typical ones, is a helpful way to raise learners' awareness of collocability. Because the options in the MC questions are semantically similar, participants would have difficulty in choosing unless they have adequate understanding of the range of collocations of the base word.

In terms of gap-filling questions, these have been widely used to measure controlled productive knowledge. The rationale for using gap-filling questions to measure the learners' controlled productive knowledge of collocations is twofold: 1) gap-filling questions to a certain degree resembles real-life communication situations where the learner needs to retrieve words or collocations in response to the given contextual clues (Laufer, 1998); and 2) statistically, gap-filling questions and selected response-type items (i.e., MC items) are found to be fairly reliable measures of collocational knowledge (Bonk, 2001). Also, these types of test items are easy to construct, validate and score. An example of the gap-filling questions in the tests is as follows:

The meeting aims to address a w\_\_\_\_\_ (adj.) range of issues, from the environment to the educational system.

The initial letter of the target collocate was given, to constrain the choices of possible words and direct the learner's attention to the target collocation.

---

<sup>7</sup> Following Siyanova and Schmitt (2008), all the incorrect options in MC items meet the following criteria: 1) occurring fewer than 5 times in the 100 million words BYU-BNC: British National Corpus; and 2) having a mutual information (MI) score lower than 3.



### 3.5.2.2 Writing assignments

The collocation tests elicited data on the learner's receptive and controlled productive collocational knowledge. Free productive collocational knowledge was examined through participants' actual use of lexical collocations in written production. In order to minimize any disturbances the study might have caused to the curriculum, the participants' written production was conveniently sampled from their assignments for required writing classes, instead of having them write another essay for the sole purpose of this study. The genre and topic of writing assignments varied with the writing class the participant was in. I was unable to impose a stringent control on the writing assignments sampled for investigation, because only partial access was allowed. Some participants and instructors of the writing classes were less enthusiastic about providing writing samples for research purposes, and others could not do so because they had different tasks in the writing class. The number of writing assignments collected for investigation was therefore smaller than expected. As noted above, the requirements of the assignments varied, depending on the writing class to which the participant had been assigned. The titles and genres of the assignments collected are as follows:

**Table 3-1** *The titles and genres of the writing assignments*

Exp. group	Pre (N=56)	<i>My ideal room</i> (expository: descriptive) <i>My sister's room</i> (expository: descriptive)
	Post (N= 55)	<i>My best friend and I</i> (expository: comparison) <i>Procrastination</i> (expository: problem-solution)
Control group	Pre (N=44)	<i>Should university students be required to attend classes</i> (argumentative)?
	Post (N=46)	<i>Should the amount of homework be limited</i> (argumentative)?

As seen in the table above, assignments given to the control group targeted at the genre of argumentation, while those given to the experimental group mainly focused on exposition, with slight differences in subtype (i.e., description, comparison and problem-solution). The number and type of collocations used in writing may vary with writing foci, because collocates tend to cluster within certain semantic fields (Carter, 1998; Stubbs, 1995a). It would be biased to make any between-group comparisons of collocations produced in writing assignments. The changes in free

productive knowledge were thus limited to within-group comparisons only.

### 3.5.2.3 Collocation learning worksheets

In addition to the measurements before and after the (non)intervention period (the pretest, post-test and two sets of writing assignments), the worksheets distributed to the experimental participants in the intervention sessions were collected to examine the progress of collocation learning (see Appendix 9 for the worksheets). In contrast to the measurements above, which were a form of *summative evaluation*, learner performance on the worksheets served as a *formative evaluation* of corpus-assisted collocation learning.

The 50 items in the collocation tests (see 3.5.2.1) were evenly divided into five sets of 10 items, and randomly assigned to five worksheets, each used for one weekly intervention session. In these sessions, the experimental group was instructed to complete the worksheets with the aid of six designated web-based concordancers (see the intervention sessions in 3.4). The items were approximately the same as those in the pretest except for some additions of options in MC questions and gaps in gap-filling questions, to encourage exposure to more collocations of the same base word while consulting corpora. For example:

Pretest: The professor \_\_\_\_\_ a speech about current economic situations.  
(Choose one INCORRECT option)  
(A) sent (B) delivered (C) gave

Worksheet: The professor \_\_\_\_\_ a speech about current economic situations.  
(Choose one INCORRECT option)  
(A) delivered (B) gave (C) made (D) talked

Generally, one to three collocations of a base word were assigned for corpus consultation. It is characteristic of concordance-based materials to have an intensive focus on a small number of target lexical items and/or collocations (Thurstun and Candlin, 1998). The idea behind DALC in this study was not to expand the learner's vocabulary size by presenting new words: rather, it aimed to build on his/her existing knowledge of base words (presumably knowledge of word meanings) and to sensitize the learner to the range of typical collocations of these words through an intensive exposure to and self-induced analysis of corpus language data.

### 3.5.3 Analyzing data on collocational knowledge

The previous section (3.5.2) has outlined the instruments for eliciting data relating to learners' collocational knowledge. This section goes on to detail the methods of analyzing collocation tests, worksheets and writing assignments.

#### 3.5.3.1 Marking tests and worksheets

The collocation tests and worksheets were manually marked by myself. Both instruments contained two types of questions, MC questions and gap-filling questions. Marking the MC questions was rather straightforward, since there was only one correct answer to each question. One point was given if the answer was correct, while no point was gained for incorrect answers. On the other hand, marking gap-filling questions was less straightforward, because of a wider range of possible answers, though the first letter of each collocate was provided to limit the scope of possible answers (see 3.5.2.1). The gap-filling questions were marked with reference to 1) *The BBI Dictionary of English Word Combinations* and *Oxford Collocations Dictionary for Students of English*; 2) BYU-BNC: British National Corpus (100 million words); and 3) native speaker intuition if necessary. The criteria for judging the acceptability of the answer to a gap-filling question are as follows:

**Table 3-2 The criteria for marking gap-filling questions**

Description of possible answer	Acceptability
The answer matches the desired response.	acceptable
The answer matches the desired response but is incorrectly inflected. Although it is ungrammatical, the correct choice of collocate demonstrates knowledge of that particular collocation. For example: No one was injured Tuesday when a school bus <u>*catch/*catches/*catching</u> fire.	acceptable
The answer does not match the desired response, but it is contextually appropriate and the combination is accepted in any one of the reference books or corpus ( $\geq 5$ occurrences).	acceptable
The answer neither matches the desired response nor fits into the given context, but the combination is accepted in any one of the reference books or corpus ( $\geq 5$ occurrences). For instance: To highlight the importance of energy-saving, teachers should <u>*give</u> an example for students by having lights switched off whenever they leave the room. The collocation <i>give an example</i> is listed in the reference books, but does not lend itself to the given context, so it is regarded as unacceptable.	unacceptable
The answer neither matches the desired response nor fits into the given context, and the combination is not accepted in any one of the reference books or corpus.	unacceptable
The answer does not match the desired response but is contextually	seek NS intuition



appropriate. The combination is not accepted in any one of the reference books or corpus.	for acceptability
No answer	unacceptable

All the answers to the gap-filling questions were marked as either *acceptable* or *unacceptable* based on the set of criteria above. Answers judged to be acceptable were awarded one point, while unacceptable answers gained no point. As for NS judgment, acceptability was determined on two out of three NS responses.

### 3.5.3.2 Analyzing tests and worksheets

The pretest and post-test were administered to the experimental group and control group, so four sets of test scores (i.e., 2 tests x 2 groups) were gathered and subjected to statistical analyses, using SPSS 13 (Statistical Package for the Social Sciences) for within-group and between-group comparisons. *Descriptive statistics*, including the mean or median, and standard deviation (SD), were calculated to examine the participants' performance on the pretest, post-test and worksheets. Descriptive statistics were used to 'characterize or describe a set of numbers in terms of central tendency and to show how the numbers disperse, or vary, around the centre' (Brown and Rodgers, 2002:122). *Inferential statistics*, on the other hand, were used to make comparisons within and between groups:

1) As regards entry-level collocational knowledge, the pretest scores were compared between groups using the *Mann Whitney U test*<sup>8</sup> to see if the two groups were closely matched in terms of receptive and controlled productive knowledge of collocations before the intervention.

2) In order to measure the changes in collocational knowledge brought about by the (non)intervention, the scores of the pretest and post-test were compared within groups using the *Wilcoxon signed-rank test*<sup>9</sup> (experimental group) or *paired sample t-test*<sup>10</sup> (control group).

<sup>8</sup> The Mann-Whitney *U* test was performed because the pretest scores of the experimental group were not normally distributed: the Kolmogorov-Smirnov test (K-S test)  $D(97)=0.11$ ,  $p = .010$ .

<sup>9</sup> The Wilcoxon sign-ranked test was performed because the test scores of the experimental group were not normally distributed,  $D(97)=0.11$ ,  $p = .010$ ,  $D(97)=0.11$ ,  $p=.006$ .

<sup>10</sup> The paired sample *t*-test was performed because the test scores of the control group had normal distributions,  $D(76)=0.08$ ,  $p = .200$ ,  $D(76)=0.10$ ,  $p=.084$ .

3) To probe further into the effects of DALC on the experimental participants' receptive and controlled productive knowledge of collocations, the *Wilcoxon signed-rank test*<sup>11</sup> was performed to see if there were any significant differences at these two levels of collocational knowledge before and after the intervention.

### 3.5.3.3 Identifying collocations in writing assignments

Free productive collocational knowledge was operationalized as the collocation production in writing assignments. The lexical collocations used in the assignments were identified manually by myself, with reference to Benson, Benson and Ilson's (1997) classification of lexical collocations (see 2.2.3.2, Table 2-3). While this classification served as a good starting point for identifying lexical collocations, it required a slight modification to encompass a broader range of lexical collocations produced by learners in this study. For example, *hurt one's feelings* was a common VN collocation, but did not fit into either of the two subtypes of VN collocations in Benson, Benson and Ilson (ibid.). The varieties of VN collocations seem to be wider than Benson, Benson and Ilson (ibid.) had prescribed. Moreover, a body of empirical evidence has shown that VN collocation has posed the greatest difficulty for SL/FL learners, compared with other types of lexical collocations (Bahns, 1993; Chang, 1997; Nesselhauf, 2003). Given the difficulty VN collocations cause to learners, I wished to probe further into learners' written production of this particular type of lexical collocations. The study therefore did not confine itself to the two subtypes of VN collocations noted above, but took into account all the VN collocations produced in the sampled writings. Also, it did not further distinguish between CA and EN collocations under the broad category of VN collocations as Benson, Benson and Ilson (ibid.) did, because this classification may preclude a number of VN collocations used by the participants.

In addition to the modifications on the category of VN lexical collocations, several grammatical collocations in Benson, Benson and Ilson (ibid.) were also included for investigation in the present study if they contained verbs or adjectives that modified nouns. For example, *the book is too hard to read*, this construction consisted of a

---

<sup>11</sup> The Wilcoxon sign-ranked test was performed because all the scores for the MC and gap-filling questions did not have normal distributions,  $D(97)=0.15, p=.000$ ,  $D(97)=0.13, p=.000$ ,  $D(97)=0.09, p=.042$ ,  $D(97)=0.12, p=.002$ .

predicate adjective and a following infinitive, so it was classified as a grammatical collocation in Benson, Benson and Ilson (ibid.), whereas in this study two lexical collocations would be extracted from this construction: *the hard book* (adjective-noun lexical collocation) and *read the book* (VN lexical collocation). Although such utterances were constructed in grammatical patterns, there were nonetheless lexical collocational relations among the constituents: *hard* being the adjective collocate and *read* being the verb collocate of the noun *book*. Hence, for the purposes of assessing learners' knowledge of lexical collocability, grammatical constructions entailing lexical collocational relations among constituents were also counted as lexical collocations in the present study. The revised working classification of lexical collocations is in Table 3-3.

**Table 3-3 The working classification of lexical collocations**

Type	Example
verb + noun	<i>read the book</i> : the book to read; the book that should be read; the book is too hard to read (also see adjective + noun).
	<i>write a letter</i> : he wrote her a letter; he wrote a letter to her.
	<i>adhere to the rule</i> : verb + preposition + noun
	<i>unload the car</i> : He watched them unload the car.
	<i>dye hair</i> : She dyed her hair red. (also see adjective + noun )
adjective + noun	<i>hard book</i> : The book is too hard to read (also see verb + noun) <i>red hair</i> : She dyed her hair red. (also see verb + noun )
noun + verb	<i>bees buzz</i>
noun of noun	<i>a pack of dogs</i>
adverb + adjective	<i>deeply moved</i>
verb + adverb	<i>argue heatedly</i>

In summary, Benson, Benson and Ilson's (ibid.) classification of lexical collocations was revised to encompass a wider range of lexical collocations produced in the participants' writing assignments. The lexical collocations identified based on the working classification above were then coded and subjected to statistical analyses, as detailed in the next subsection.

#### **3.5.3.4 Coding and analyzing collocations in writing assignments**

The lexical collocations identified in the participants' writing assignments were evaluated in terms of *number*, *type* and *acceptability*. Also, the *miscollocations* produced by the experimental participants were examined. As noted earlier (see 3.5.2.2), the differences in writing topic and genre may lead to different profiles of



collocation production, so the comparisons of free collocation production were limited to within-group ones only.

1) The *number* of lexical collocations in the writing assignments was tallied manually by myself. As the four sets of writing samples (2 groups x 2 points in time) varied in length, the *number of collocations per 100 words* was used for comparisons, rather than the *raw number of collocations in a piece of writing*. A *paired sample t-test* was performed to compare the numbers of collocations per 100 words, because these four sets of data were all normally distributed<sup>12</sup>.

2) In terms of the *types* of collocation, following Howarth's (1998a) continuum of lexical composites (see 2.2.3.2) and Nesselhauf's (2003:225) classification of word combinations<sup>13</sup>, the lexical collocations in the writing assignments were classified into three types along a cline of restrictedness: *free combinations (FCs)*, namely the 'combinations in which a possible restriction on the substitutability of elements is due to their semantic properties', for example, *white curtain*, *need something*; the second type was *restricted collocations (RCs)*, the 'combinations in which this restriction is to some degree arbitrary', for instance, *spend time*, *give somebody a hand*; and the latter type was *fixed expressions (FEs)* that were highly restricted, and did not allow any substitution of the constituents, for example, *jump the gun* (see 2.2.3.1 for sets of criteria for classifying collocations).

While coding FCs and FEs was rather straightforward, that of RCs was much less so. Therefore, more specific criteria were set out to identify RCs. A word combination would be classified as a RC if at least one of the following criteria (adapted from Nesselhauf, 2003) applied:

1. The sense of the node (collocate) is so specific that it only allows its combination with a small set of collocates (nodes).
2. The node (collocate) cannot be used in this sense with all collocates (nodes) that are syntactically and semantically possible.

For example, the word combination *need something* is a FC, as the verb *need* can be combined with a great number of nouns, whereas *spend time* would be considered a RC, as the verb *spend* is used in a restricted sense and it cannot be substituted with

<sup>12</sup>  $D(53) = 0.07, p = .200, D(53) = 0.08, p = .200, D(44) = 0.11, p = .200, D(44) = 0.12, p = .118$ .

<sup>13</sup> Nesselhauf (2003) distinguishes between 3 major types of word combinations: free combinations, collocations and idioms.

other semantically possible verbs when used with *time* (e.g., *\*to pay time*). Three dictionaries were also used to determine if a word was used in a restricted sense: the *Collins COBUILD English Dictionary*, the *Macmillan English Dictionary for Advanced Learners of American English* and the *Oxford Advanced Learner's Dictionary*. Note that the term *collocation* is used here interchangeably to denote FCs and RCs, unless indicated otherwise. The coding was conducted by myself three times with two-week intervals (in the weeks of 11<sup>th</sup> January, 25<sup>th</sup> January and 8<sup>th</sup> February 2010) to ensure consistency. The Fleiss's Kappa statistic<sup>14</sup> was computed to determine the intra-rater reliability among the three coding sessions: the reliability was found to be 0.71 ( $p < .001$ ). A Kappa value between 0.61 – 0.80 indicates substantial agreement among coding sessions (Fleiss, 1981).

In order to understand the changes in the types of collocation produced before and after the (non)intervention period, the proportions of the types of collocation were compared using the *Wilcoxon signed-rank test* and *paired sample t-test*. The *Wilcoxon signed-rank test* was performed to compare the proportions of FCs and RCs produced by the experimental group, and those of FCs produced by the control group, because most of them were not normally distributed<sup>15</sup>. In contrast, a *paired sample t-test* was performed to compare the proportions of RCs in the control group's writings, as the data had normal distributions<sup>16</sup>.

3) The *acceptability*<sup>17</sup> of the lexical collocations used in the writing assignments was judged against both corpus frequency counts (BYU-BNC: British National Corpus) and human judgment. As Howarth (1998a:29) points out,

[i]t must be recognized that decisions about the acceptability of combinations that occur individually at very low frequencies must continue to rely heavily on human judgment. The absence of a possible combination from dictionaries and even large corpora cannot reasonably exclude it from consideration. Additionally, the collocations of most interest in studying acquisition are not typically fixed enough for automatic identification.

<sup>14</sup> The Fleiss's Kappa statistic measures the degree of agreement among multiple raters (more than two) or rating sessions on assigning categorical ratings.

<sup>15</sup> K-S test:  $D(53) = 0.19, p = .000, D(53) = 0.13, p = .019, D(53) = 0.19, p = .000, D(53) = 0.15, p = .007, D(44) = 0.15, p = .019, D(44) = 0.06, p = .200$ .

<sup>16</sup> K-S test:  $D(44) = 0.11, p = .197, D(44) = 0.06, p = .200$ .

<sup>17</sup> Following Howarth (1998a) and Nesselhauf (2003), the term *acceptability* was adopted to denote the degree to which a word combination conforms to the typical usage in the target language, taking into account the context in which it is embedded. This term is more suitable than others used in previous studies (e.g., *commonness*) to describe FL learner production which is often relatively uncommon in the TL.

Following Nesselhauf (2003) and Siyanova and Schmitt (2008), the threshold for the *acceptability* of a collocation was  $\geq 5$  occurrences in the 100 million words BYU-BNC: British National Corpus. The mutual information (MI) score was not used, because the vast majority of learner collocations was rather free word combinations that might not have high MI scores but made perfect sense in student writings. For example, *a perfect room*, *the original wall*, such word combinations could hardly be labelled as *unacceptable* given the contexts in which they were embedded. In this sense, a follow-up human judgment that took into account the context in which a combination occurred was necessary: the acceptability of the combinations that had fewer than five occurrences in BYU-BNC was judged again manually, with reference to the extended contexts in which they occurred. I judged the acceptability of these potential unacceptable collocations three times with two-week intervals (in the weeks of 5<sup>th</sup> July, 19<sup>th</sup> July and 2<sup>nd</sup> August 2010) to ensure consistency. After my coding, some combinations remained problematic as to their typicality/commonness as a collocation and their acceptability in the contexts, so three native speakers of English were invited to judge the acceptability of such combinations. Inter-rater reliability refers to, as Cohen, Manion and Morrison (2007:119) put it, ‘whether another observer with the same theoretical framework and observing the same phenomena would have interpreted them in the same way’. A problematic combination would be accepted if two out of three NS raters judged it acceptable. Unfortunately, when it came to the acceptability of FL learner collocation production, judgment seemed to vary considerably among NS raters<sup>18</sup>. The inter-rater reliability was 0.13,  $p < .05$  (Fleiss’s Kappa), indicating only slight agreement among three NS raters. The resulting rates of acceptability were compared using the *Wilcoxon signed-rank test* because of the non-normal distributions<sup>19</sup>.

4) The *miscollocations* produced by the experimental group were classified based on the possible contributing factors, to illuminate the qualitative changes that occurred in the participants’ productive knowledge of collocations over the intervention. In this study, the classification of miscollocations was predominantly data-driven, and

---

<sup>18</sup> However, the high variability in judging the acceptability of learner collocations may well be attributable to the *dichotomous* rating system, as the acceptability may be better represented in a *continuous* manner. The limitation of this rating system is discussed in Chapter 7(7.2.1).

<sup>19</sup> K-S test:  $D(53) = 0.16, p = .001, D(53) = 0.17, p = .001, D(44) = 0.15, p = .010, D(44) = 0.12, p = .093$ .



developed in the light of the instances found in the present dataset. I examined the data and developed the coding scheme accordingly:

*Table 3-4 The coding scheme for miscollocations*

Interlingual miscollocation	Literal translation
	Deviant translation
Intralingual miscollocation	Using synonyms
	Analogy
	Paraphrase
	Using de-lexicalized words
	Synformy
	Grammatical errors
Unintelligible combination	

Based on the coding scheme shown above (see 4.2.4 for elaboration on the types of miscollocation), two Chinese/English bilingual raters and myself coded the miscollocations independently to strengthen consistency. Inter-rater reliability (Fleiss’s Kappa) was found to be 0.69 ( $p < .001$ ), indicating substantial agreement among raters. Finally, the number and proportion of each type of miscollocation were tallied.

### 3.6 Collecting and analyzing data on thinking processes

This section outlines the data collection and analysis methods employed to address the second research question:

RQ2. What is the nature of the thinking processes EFL learners engage in during the data-driven approach to learning collocations?

As the data on the thinking processes underlying corpus consultation could not be obtained directly from observation, *mentalist* and *behaviourist* measures were taken to tap into the learner’s mind as s/he engaged in a DALC task.

#### 3.6.1 Collecting data on thinking processes during corpus consultation

The *mentalist* measure used was concomitant think-aloud, which gathered the learner’s concurrent verbal reports of his/her thinking processes on-task. Think-aloud verbal reports uncover what lies beneath the corpus consultation behaviour, including the ongoing cognitive processes and language learning strategies employed in the

course of a language task (Woodfield, 2010). In addition, I also collected *behaviouristic* data in parallel with the think-aloud verbal reports, namely the corpus resources consulted and the queries entered as the learner articulated his/her thoughts, to complement the *mentalistic* data in reconstructing the thinking processes during corpus consultation.

### 3.6.1.1 Rationale for think-aloud

Think-aloud is a verbal data elicitation technique in which a respondent is asked to speak out whatever thoughts cross his/her mind as s/he undertakes a problem-solving task (Bowles, 2010b; Ericsson and Simon, 1993). Think-aloud explores the otherwise hidden cognitive processes underlying the observable behaviour of the learner (Van Someren, Barnard and Sandberg, 1994). Cohen (1996:7) regards think-aloud as a self-revelational verbal report, or a 'stream-of-consciousness disclosure of thought processes while the information is being attended to'. In this study, think-aloud verbal reports allowed me to gather real-time information about the ways in which the learner approached the DALC task, including what his/her linguistic hypothesis was, how it was tested with corpus resources, what strategies s/he employed to induce target collocations from concordances, and so on. As Grenfell and Harris (1999:54) aptly describe, 'it is not easy to get inside the "black box" of the human brain and find out what is going on there. We work with what we can get, which, despite the limitations, provides food for thought'.

The major strength of think-aloud lies in its immediacy. Unlike other elicitation techniques that collect data at the closure of an activity, think-aloud gathers verbal data in real-time as the informant performs a language task. It thus reduces the threats of memory failure or distortion that result from time lapse, as is the case with retrospection methods (Wade, 1990). Ericsson and Simon (1993) prefer think-aloud to other types of verbal report (e.g., self-report questionnaires, self-observational diary/journal or retrospection) for its immediacy in terms of generating data and placing less strain on memory in reconstructing the thought processes. Immediacy was particularly important in the present study, because in the DALC task the respondents needed to process a significant amount of corpus data before they could generalize target collocations. It was thus likely that they would lose track of the

details in the thought processes after such a painstaking endeavour. Hence, concomitant think-aloud was used to elicit the participants' thought processes concurrently as they performed the DALC task, instead of gathering data in retrospection (e.g., stimulated recall<sup>20</sup>).

Another advantage of using think-aloud is that while some elicitation techniques raise concerns as to *halo effect*, namely the respondent attempting to please the researcher by giving responses that they think are desired (Mackey and Gass, 2005), think-aloud engages respondents in multitasking (i.e., problem-solving and verbalizing thought processes simultaneously), so they are less likely to tailor the responses to meet the researcher's expectation. As Young (2005:22) suggests,

during the think-aloud process, when a participant is absorbed in a given activity, the completion of this task will take precedence thus limiting the available cognitive 'space' for the participant to formalize desirable thoughts.

Within the field of SLA, think-aloud has been used extensively in studies on learner strategies (cf. Bowles, 2010a), particularly reading strategies, but few studies have employed think-aloud in the research on technology-enhanced learning, with the exception of Sun (2003) (as discussed in 2.5.2). Given the revealing and instantaneous nature of think-aloud verbal reporting, the present study employed this method to capture the ongoing thinking processes as the respondent consulted corpora for target collocations. This study also complemented verbal reports with parallel corpus query records, so as to reconstruct the thinking processes as completely as possible.

### 3.6.1.2 Validity issues of think-aloud

Despite the strengths of think-aloud noted above, there has been concern in the literature regarding the *reactive effects* or *reactivity* of concurrent verbal reporting: that is, whether concomitant verbalization *per se* alters thought processes (cf. Bowles and Leow, 2005). Nunan (1992) raises doubts about the extent to which verbalization truthfully reflects cognitive processes, since the act of verbalizing may alter the processes. If this is the case, the change(s) in the thought processes caused by concurrent verbalization would cast doubt on the validity of the verbal reports. However, contrary to such common concern, a body of empirical evidence has shown

---

<sup>20</sup> See Gass and Mackey (2000) for a detailed discussion of stimulated recall.



that concurrent verbalization does not have a significant impact on the outcomes of the tasks given. For example, Leow and Morgan-Short (2004) measured FL learners' knowledge of the impersonal imperative in Spanish, in terms of reading comprehension, intake and controlled written production. They found no statistically significant differences in the above three respects between the think-aloud participants and their non-think-aloud counterparts. Likewise, Medina (2008) also found that thinking aloud had neither positive nor negative effects on L2 learning (as measured by a recognition and written production test). Her findings demonstrated that though concurrent verbalization delayed task completion time, but not insofar as to have a substantial impact on the learning outcomes. Notwithstanding the empirical evidence above, reactive effects may be linked to task types and individual differences (cf. Stratman and Hamp-Lyons, 1994), so it would be unwise to overgeneralize the empirical findings and to assume that the present investigation was impervious to the threat of reactivity. While no inquiry has been conducted to investigate the reactive effects on concurrent verbal reports during corpus consultation, the method was employed with caution in this study, bearing in mind the methodological limitations and the immense complexity of human cognition.

Moreover, the limited attentional resources that humans have available for multitasking (e.g., verbalizing thoughts and performing a task simultaneously) may render concurrent verbal reports incomplete and fragmented (Færch and Kasper, 1987). To compound the problem, the quality of verbal reports depends largely on the verbal facility of the respondent (Van Someren, Barnard and Sandberg, 1994): a respondent who is not verbally competent or comfortable enough may not be able to verbalize thoughts fully. Despite such limitations, think-aloud has hitherto been a relatively more adequate measure to tap into ongoing thought processes, compared with other types of elicitation method. As Ericsson and Simon (1987:51) put it, '[a]lthough they are not fully adequate for catching the fine grain of thought processes, verbal protocols ... have provided data at the highest densities we have as yet attained'. To echo Ericsson and Simon's (ibid.) comment, Cohen (1998:38-39) states:

[w]hereas the reliability of mentalistic measures has been questioned in comparison with behaviouristic measures, research has demonstrated that verbal reports, elicited with care and interpreted with full understanding of the circumstances under which they were obtained, are, in fact, a valuable and a thoroughly reliable source of information about cognitive processes.

Bearing in mind the limitations of think-aloud, I endeavoured to gather data on thinking processes with caution and guard against potential threats of 'under reporting' (Matsumoto, 1994) on the part of the respondent by complementing verbal reports with parallel corpus queries when interpreting results, and threats of 'over reporting' (Jourdenais, 2001) by refraining from cueing a particular response or asking leading questions.

### 3.6.1.3 Piloting think-aloud

The think-aloud method was piloted to determine if it was appropriate as a means for eliciting cognitive processes during corpus consultation. Three volunteers (undergraduate EFL learners) participated in the piloting think-aloud session on 1<sup>st</sup> August 2008. Prior to this session, the participants were informed of the aim of the pilot study and given training on how to use web-based concordancers. Each participant was given one hour to do a DALC worksheet (as will be discussed in 3.6.2) and concurrent think-aloud simultaneously. The piloting sessions were video-recorded to capture the respondents' verbal reports and paralinguistic information (e.g., facial expression, body language). Their parallel corpus queries were also documented along with verbal protocols.

The purpose of piloting think-aloud was to understand the extent to which it was viable as a measure to gather data on thinking processes during corpus consultation. Unlike piloting *data-elicitation instruments* (e.g., test or questionnaire), piloting *data-elicitation methods* such as think-aloud relies predominantly on the researcher's judgment of whether such a method is fit for the research purposes. The following discussion is based on my evaluation of the viability of think-aloud as a data-elicitation measure in this study. My piloting experience showed that the combination of mentalistic data (i.e., think-aloud verbal protocols) and behaviouristic data (i.e., corpus queries) usefully revealed learners' thought processes as they undertook the DALC task. Verbal reports along with parallel corpus queries allowed me to follow the trails of learners' thinking, thereby reconstructing the processes in which corpus data was cognitively processed by learners' mental capacities. On the other hand, think-aloud, unlike questionnaire or interview, was not a common data-elicitation measure with which the participants might be familiar. Therefore, the

researcher's verbal instructions were necessary but insufficient: in some instances the participants were uncertain about what to report or tended to pause to organize thoughts before articulating, which resulted in lapses between the time of actual cognizing and that of verbalizing. The piloting experience had implications for the main study to complement verbal instructions with a visual demonstration (e.g., a modelling video clip) whereby the respondents could observe how think-aloud was carried out. Also, it was found that paralinguistic information was scant in this particular task: the respondents were required to solve the problems individually with computers and web-based concordancers, so little body language was found. Therefore, audio-recording instead of video-recording would suffice to gather data in the main study.

#### **3.6.1.4 Conducting think-aloud in the main study**

The think-aloud sessions in the main study were conducted in April and May 2009 after the DALC intervention period (see the intervention in 3.4). 17 participants from the experimental group volunteered to participate. They were informed of the purpose and procedure of the think-aloud interviews before giving their consent for participation (see the consent form in Appendix 3). Prior to the think-aloud interviews, the participants were given a modelling session where they watched a video clip demonstrating how to think aloud. Ericsson and Simon (1993) suggest that the training of respondents in providing verbal reports ensures consistency among respondents, and improves the quality of data. The think-aloud sessions were conducted individually with myself, each taking approximately an hour. In the think-aloud session, the participant was asked to do a DALC task (the worksheet used for think-aloud sessions will be discussed in 3.6.2) with the aid of the six designated web-based concordancers (see 3.4 for the concordancers) or any resources available online. Meanwhile, the participant was required to verbalize his/her thought processes concurrently while undertaking the DALC task. As the purpose of the think-aloud interview was to elicit the ongoing mental processes during corpus consultation, rather than the learning outcomes of such endeavours, the respondents were instructed to focus on verbalizing the processes as fully as possible and not to worry about their performance on the worksheet. Regarding the language for verbal reporting, Cohen (1998) suggests that the choice should be left to the respondent, since the use of the



TL may be at the expense of generating adequate data. The present study left open the choice of language for reporting to the respondents. All the respondents carried out the task predominantly in their L1, Mandarin Chinese, with sporadic use of English when referring to the words on the worksheet or concordances. Note that in the think-aloud sessions, the respondents were allowed to use any resources available online in addition to the designated web-based concordancers, so that the consultation endeavours may approximate real-life ones as closely as possible. With the respondents' consent, the think-aloud verbal reports were audio-recorded, and the corpus queries were manually documented by myself. To minimize any interference in the respondents' thinking processes, I did not engage in the task, but only provided occasional technical assistance as necessary, or prompted them to speak if pauses in verbalization persisted.

### **3.6.2 The worksheet for think-aloud**

The collocation learning worksheet in the think-aloud sessions (see Appendix 10) was not used as a data-elicitation instrument but a prompt to trigger verbal reports on thinking processes during corpus consultation. That is, the answers to the questions on the worksheet *per se* were not analyzed, rather, the thinking processes verbalized (mentalistic data) in the course of the task and the parallel corpus queries (behaviouristic data) were gathered and subjected to thematic analysis. The worksheet consisted of six question types: contextualized MC questions, de-contextualized MC questions, gap-filling questions, miscollocation identification and correction, semantic grid and translation. The base words of the collocations in these questions were drawn from the 'Academic Word List' compiled by Coxhead (2000). The same process as that in developing the pretest and post-test (see 3.5.2.1) was carried out to select collocations and construct items in the think-aloud worksheet except for the validation procedure.

### **3.6.3 Analyzing think-aloud protocols and parallel corpus queries**

As noted earlier (3.6.1.4), 17 individual think-aloud sessions were audio-recorded, each taking approximately one hour. The verbal protocols were predominantly in Mandarin Chinese, and were transcribed manually by myself. Then the parallel corpus queries were aligned with the transcribed protocols.

The qualitative data analysis software NVivo was used to assist in the coding of the transcribed protocols. As noted in Chapter 2, O'Malley and Chamot's (1990) taxonomy of LLSs and Sun's (2003) cognitive skills used for concordancing (see 2.4.1.2 and 2.5.2 respectively) served as a starting point for analyzing data on thinking processes underlying corpus consultation. The rationale for basing the analysis on the former was that the LLSs therein were fairly generic, compared with specific ones such as reading strategies or vocabulary learning strategies, and therefore provided a reasonable preliminary framework for the present investigation. On the other hand, Sun's (ibid.) cognitive skills during concordancing were based on empirical findings, and thus had direct bearing on this study. A top-down approach to data analysis is strongly recommended by Ericsson and Simon (1987:29) as they argue '[a]n *a priori* analysis of the possible thought sequences generating an answer to a task is essential. Such a task analysis often reveals that an answer can be generated by several alternative processing sequences'. However, they also stress that the initial coding scheme is provisional, and needs to be modified as new findings or unanticipated themes emerge from data. Hence, the present study employed an integrated approach to protocol analysis, which set out with a predefined coding scheme noted above, but was also open to modification in the light of new findings that emerge from data.

In addition to myself, two colleagues (both Chinese/English bilinguals) were invited to code the verbal protocols to ensure the reliability of the analysis. Each coder independently coded parts of the verbal protocols with reference to parallel corpus queries. Approximately 20-30% of all the verbal protocols was subjected to inter-rater reliability check. The reliability was found to be 0.89,  $p < .05$  (Fleiss's Kappa), indicating substantial agreement among coders.

Finally, the segments selected for discussion were translated into English, the language for reporting. The translated segments were back-translated into Mandarin Chinese by the colleagues above to ensure accuracy of the translation.

### **3.7 Collecting and analyzing data on learner perceptions**

In addition to the *learning product* and *processes*, the present study also looked into *learner perceptions* of DALC. This section discusses the data collection and analysis methods employed to address the third research question:

RQ3. How do EFL learners perceive the data-driven approach to learning collocations?

A questionnaire was administered to elicit learner perceptions of DALC. The following subsections detail the development and administration of the questionnaire.

### **3.7.1 Collecting data on learner perceptions**

A questionnaire was administered to the experimental group at the closure of the intervention period on 2<sup>nd</sup> April 2009 (see Appendix 13 for the questionnaire). The questionnaire aimed to gather data on the experimental participants' previous experience of learning vocabulary, awareness of collocations, attitudes and opinions of the corpus-assisted learning experience.

### **3.7.2 The questionnaire eliciting learner perceptions**

The questionnaire eliciting learner perceptions of DALC was composed of five sections, with a total of 50 items. The questionnaire was written in the participants' L1, Mandarin Chinese, to avoid the situation where a low level of English proficiency constrained the informant's response (Mackey and Gass, 2005). The following subsections recount the development and validation of the questionnaire.

#### **3.7.2.1 Developing the questionnaire**

The questionnaire consisted of five sections.

**Section 1** contained four *demographic questions* (Items 1-4) eliciting personal information relating to gender, nationality, native language and the length of studying English as a foreign language.

**Section 2** consisted of seven *knowledge questions* (Items 5-11) used to 'tap into facts, knowledge, and information about language teaching and learning process' (Brown, 2001:31). This section asked for information on the English learning resources available to the respondent, including access to computers, the Internet, web-based concordancers and collocation dictionaries, in the format of alternative-answer questions that provided alternative answers from which the respondent must select. For example:



Do you have access to the Internet?

Yes ☐

No ☐

While the previous two sections targeted at the background information of the respondent, Sections 3-5 elicited their views or perceptions with *opinion/value questions*. Opinion/value questions, according to Brown (2001:31), 'explore the respondents' thoughts, reactions, impressions, attitudes, and outlook on various aspects of language or language learning processes'.

Section 3 gathered information on the respondent's vocabulary learning experience. There were ten items in this section, including eight *Likert-scale items* (Items 12-19), a *rank-ordered item* (Item 20) and a *check-list item* (Item 21). Likert-scale items were scored using a 6-point scale with the score of 6 for *strongly agree*; 5 for *agree*; 4 for *toward agree*; 3 for *toward disagree*; 2 for *disagree* and 1 for *strongly disagree*. Unlike the common 5-point Likert-scale items, the 6-point scale ones prevented the respondent from using the mid-point (e.g., 3 on a 5-point scale) to avoid making a real choice (Dörnyei, 2003). Neither mid-point option, nor non-substantive option such as 'I don't know' or 'no opinion', were given in this questionnaire in order to preclude responses that reflected ambivalent attitudes, a social desirability effect or concealed genuine attitudes (Smith, 1984). Item 20 was a rank-ordered question, which asked the respondent to rank order six components of word knowledge on the relative importance perceived: the rank of 1 being the most important and the rank of 6 being the least important. The six components of word knowledge were adapted from Nation (2001), including written form (orthography/spelling), spoken form (pronunciation), meaning (form-meaning link), connotation, collocation and grammatical function (see 2.2.1 for word knowledge framework). This item aimed to explore how the respondent perceived the importance of collocation relative to other components of word knowledge. Item 21 was a check-list question, which provided a list from which the respondents could check as many options as applicable to their perceptions or opinions (Brown, 2001). Building on the components of word knowledge listed in the previous question, Item 21 went on to ask the respondent to select as many components as applicable that would benefit from corpus consultation. Although web-based concordancers were used exclusively in the present study to mediate the learning of collocations, their potential for facilitating the learning of other aspects of word knowledge was also explored. This item provided insights into

the respondents' opinions of how web-based concordancers could be exploited to assist the learning of various aspects of lexical knowledge.

Section 4 contained nine 6-point Likert-scale items (Items 22-30), generating data on collocational awareness, including the respondents' awareness of collocations, to what extent they perceived the importance of collocation in language use, how they saw collocation in future English learning, and so forth. In particular, Items 23-25 set out to elicit the respondents' perceptions about the benefits of learning collocations claimed in the literature, including enhancing *accuracy*, *fluency* and *appropriateness* in language production (see 2.3.1 for discussion in the literature).

Section 5 elicited the respondents' perceptions of DALC, for example, how they perceived the usefulness of this approach and corpus data, the accessibility of web-based concordancers, how web-based concordancers could be improved to better facilitate collocation and/or language learning, and so on. Items 31-48 were 6-point Likert-scale items and Items 49-50 were check-list questions. Also, Item 50 had a follow-up open-ended question, which asked for the reason(s) for choosing the option(s). Some items in this section were designed to elicit the respondents' perceptions about the affordances of corpus-assisted language pedagogy claimed in the literature, including the *genuineness* and *profusion* of corpus data, and *learner autonomy* (see 2.3.2 for discussion in the literature). It is worth noting that Items 41-44 were negatively phrased to prevent respondents from blind responding, as Field (2005:669) suggests, '[r]everse items are important to reduce response bias, the participants will actually have to read the items in case they are phrased the other way around'. Below is an example of the negatively phrased items:

It is time-consuming to generalize collocations from corpus data.

Accordingly, the responses to these items were reversely coded before submitting to statistical analysis (see 3.7.3 for coding questionnaire responses).

### 3.7.2.2 Piloting the questionnaire

The pilot questionnaire (see Appendix 11) was administered to 19 respondents on 24<sup>th</sup> September 2008. These respondents were closely matched to those in the main study with regard to language background, levels of English proficiency and computer

literacy. They were informed of the purpose of the pilot study, and given an introduction to DALC. In order to answer the questionnaire, the respondents needed to have a hands-on DALC experience, so they were asked to undertake a DALC task before filling out the questionnaire.

The overall reliability of the questionnaire is 0.88 (Cronbach's alpha), exclusive of the background information questions in Sections 1 and 2. According to Field (2005), Cronbach's alpha coefficient around 0.70 to 0.80 indicates a good reliability of the instrument. A closer look at each section of the questionnaire reveals that the reliability coefficient of Section 3 regarding *vocabulary learning experience* was 0.71; the reliability of Section 4 concerning *collocational awareness* was 0.85; and the reliability of Section 5 relating to *perceptions of DALC* was 0.82. The results show that all sections had at least satisfactory to good reliability. After the piloting, some items were modified in terms of wording, to improve clarity.

### 3.7.3 Analyzing questionnaires

The questionnaires completed by the experimental participants were coded<sup>21</sup> and submitted to statistical analyses. The background information questions and the Likert-scale items were subjected to descriptive statistics to calculate frequency counts or mean ratings, so as to provide an overview of the respondents' perceptions of the statements in the items. In addition, for the 6-point Likert-scale items, a score of 3.5 out of 6 served as a cut-off point to separate *positive* and *negative* attitudes: mean ratings significantly higher than the mid-point 3.5 were interpreted as a *positive* attitude toward the statements in the items, whereas those significantly lower than this point were taken as a *negative* attitude. *One sample t-test* was performed to determine whether the mean rating of an item differed from the mid-point 3.5 in a statistical sense. Furthermore, *Spearman's correlation* was performed to determine the strengths of association between the learners' performance data and perception data: that is, the correlation between the degree of progress in collocation tests and Likert-scale responses, in order to understand the link(s) between learner performance and learner perception.

---

<sup>21</sup> Items 41-44 were negatively phrased to reduce response bias, so the responses to these items were reversely coded, with a score of 1 being *strongly agree* and a score of 6 being *strongly disagree*.



The responses to the rank-ordered item, Item 20, were submitted to Friedman's ANOVA and *post hoc* Wilcoxon sign-ranked tests to determine if the mean rank of a component differed significantly from others. On the other hand, the check-list items, Items 21 and 50, were subjected to Cochran's test and *post hoc* McNemar's test to determine if an option on the given list was significantly different from other options. Item 49, another check-list item, was not submitted to inferential statistics, because the options given in this item did not need to be compared against one another: this item asked respondents to choose the concordancer feature(s) that they found important for consulting collocations.

### 3.8 Validity and reliability of the research

As noted in Section 3.1.2, the present study was methodologically guided by *positivism*, so considerable effort was devoted to strengthen validity and reliability with a positivist rigour to ensure the quality of the research.

Validity is concerned with 'the extent to which a piece of research actually investigates what the researcher purports to investigate' (Nunan, 1992:14). In this regard, this study aimed to investigate three key dimensions of DALC (i.e., the *learning outcomes, thinking processes underlying corpus consultation and learner perceptions*). Such a multidimensional inquiry may guard against threats to validity, as the potential weaknesses of individual methods could be complemented by the strengths of others (Brannen, 1992). As most previous studies on corpus-assisted language learning have only looked into performance data (see 2.5), it is imperative to look behind learner performances to explore the underlying mechanisms contributing to the claimed learning outcomes. Furthermore, as this study set out to examine the efficacy of corpus resources as language learning tools, it is essential to understand how learners themselves perceive the usefulness thereof in assisting language learning. Hence, this study looked into various dimensions of DALC with the aim of exploring the full extent to which it was adequate as a pedagogical approach. In view of this, multiple approaches were taken to tap into each dimension: as regards the learning outcomes, collocational knowledge was measured at three levels (i.e., receptive, controlled productive and free productive knowledge) to strengthen the validity of production data; learner performance was assessed *summatively* at the closure of the

intervention as well as *formatively* over the course of the interventions sessions; the cognitive processes that contributed to the learning outcomes above were reconstructed from a combination of *mentalist data* and *behaviouristic data* to ensure a truthful representation of such processes (see 3.6.1.2 for discussion on potential threats to the validity of verbal reports). Validity of the current research was strengthened through the methodological triangulation noted above.

Reliability is 'the extent to which our measurement instruments and procedures produce consistent results in a given population in different circumstances' (Dörnyei, 2007:50). Various measures were taken to ensure reliability in this research (as discussed at length in Sections 3.5-3.7). To recapitulate, the development of data elicitation instruments (e.g., the pretest, post-test, worksheets and questionnaire) followed a disciplined validation procedure, including piloting with a smaller matched sample from the same population of the main study or submitting to statistical reliability checks (see instrument validation procedures in 3.5.2 and 3.7.2). Also, as the think-aloud method has rarely been used in CALL research, it was piloted to ensure its validity as a verbal data elicitation measure and its feasibility in this particular research setting (see 3.6.1.3 for piloting think-aloud). Regarding data analysis, reliability was ensured through inter-rater or intra-rater reliability checks that were based on a set of objective criteria generated from the theoretical and empirical literature (e.g., the criteria for identifying lexical collocations, those for classifying types of collocation, those for judging the acceptability of collocations, or the taxonomy of LLSs). In addition, reference tools such as collocation dictionaries or corpora were also consulted to strengthen the reliability of data analysis (see reliability checks in data analysis in 3.5.3, 3.6.3 and 3.7.3).

### **3.9 Ethical issues**

This research was conducted in accordance with the guidelines for good practice in applied linguistics, as recommended by the British Association for Applied Linguistics (BAAL), which are summarized as follows:

1. The rights, interests, sensitivities and privacy of the participants should be respected.
2. Informed consent from the participants should be obtained prior to the research project, however, their decision to withdraw at any point in the course of the research should also be respected.
3. The participants' identities should be kept confidential and their anonymity be ensured.
4. Deception and covert research is unacceptable.

Prior to the main study, the participants were informed of the aims of the research and the details of the data collection procedure. As information on their language performance was to be gathered, the participants were reassured that the results would be used for research purposes only, and would not affect their academic achievement in the course 'General English' on which the intervention was based. A consent form (see Appendix 1 and Appendix 2) was given to the participants and the instructors in order to obtain their consent to participation, and at the same time to acknowledge their rights to withdraw at any time in the course of the research. The data elicitation instruments were anonymous, so that the participants would not have the stress of being assessed, nor feel the need to provide socially desirable responses to the questionnaire. But for the purposes of data analysis, they were asked to leave student number on the test sheets and questionnaire. Also, a small group of participants volunteered to do think-aloud, they were given another consent form prior to the think-aloud sessions (see Appendix 3). The think-aloud sessions were conducted on an individual basis, and the participants were later identified by codes when reporting to protect their real identities. All the information provided by the participants was kept confidential. As Frankfort-Nachmias and Nachmias (1992) suggest, '[t]he obligation to protect the anonymity of research participants and to keep research data confidential is all-inclusive. It should be fulfilled at all costs unless arrangements to the contrary are made with the participants in advance'.

Cohen, Manion and Morrison (2007:51) point out that the major ethical dilemma in social research is one that 'requires the researchers to strike a balance between the demands placed on them as professional scientists in pursuit of truth, and their subjects' rights and values potentially threatened by the research'. Admittedly, the major ethical concern in the present study was the rights of control group members, as they had been denied access to DALC intervention. As Fitz-Gibbon (1996) argues, it is only unethical if the researcher knows which group is to be disadvantaged. In this sense, DALC had not been proven beneficial for learning at the time of assigning participants to groups, and such group-assigning was random, so in theory the control group members were not purposefully disadvantaged. However, in practice, it is undeniable that the causal inference between DALC intervention and collocation learning was verified at the expense of the learning opportunity of the control group. I was acutely aware of this ethical concern, and endeavoured to minimize any



disruptions the research might have caused to these participants, in particular making sure that the non-intervention would not disadvantage them in terms of academic achievement. The target collocations were also taught to the control participants as an integral part of vocabulary learning in the course 'General English', only without the opportunity to consult corpora. Also, the target collocations were not tested as part of the assessment in the course, so as not to affect their grades of the course.

Cohen, Manion and Morrison (2007:56) stress the ethics of social research:

[s]ocial scientists generally have a responsibility not only to their profession in its search for knowledge and quest for truth, but also for the subjects they depend on for their work. Whatever the specific nature of their work, social researchers must take into account the effects of the research on participants, and act in such a way as to preserve their dignity as human beings. Such is ethical behaviour.

Adhering to the guidelines recommended by BAAL and the literature on the ethics of educational and social research, this study endeavoured to respect and protect the participants as well as the information they provided at all costs.

### 3.10 Summary

This chapter has provided a detailed account of the data collection and analysis methods employed to elicit data on the three key dimensions of the research: the *learning product*, *learning processes* and *learner perceptions* of DALC.

**Table 3-5 A summary of research methods**

Dimension of DALC	Data collection method/instrument	Data analysis method
Learning product (RQ1)	Pretest and post-test Writing assignments Worksheets	Descriptive statistics, Mann-Whitney test, paired sample t-test or Wilcoxon signed-rank test, Spearman's correlation.
Learning processes (RQ2)	Think-aloud Corpus queries	Thematic analysis of think-aloud verbal protocols with reference to parallel corpus queries.
Learner perceptions (RQ3)	Questionnaire	Descriptive statistics, one sample t-test, Spearman's correlation, Friedman's ANOVA and <i>post hoc</i> Wilcoxon sign-ranked test, Cochran's test and <i>post hoc</i> McNemar's test.

The learning product of DALC was measured by the changes that occurred in the learners' *receptive*, *controlled productive* and *free productive knowledge* of collocations. The elicitation instruments were collocations tests, worksheets and writing assignments. The learning processes of DALC were examined through the thinking processes underlying corpus consultation. A combination of mentalistic

(think-aloud verbal protocols) and behaviouristic (corpus queries) measures was used to shed light on such processes. The learner perceptions of DALC were elicited by means of a questionnaire. It was hoped that such a multidimensional investigation may construct a more complete picture of this innovative pedagogical approach. The findings and discussion follow in the next three chapters: Chapter 4 addresses the learning product of DALC (RQ1); Chapter 5 focuses on the learning processes of DALC (RQ2); and Chapter 6 is concerned with the learner perceptions of DALC (RQ3).

## Chapter 4 Evaluating Learning Product of DALC

This chapter reports the quantitative and qualitative changes in the participants' collocational knowledge brought about by DALC intervention. It addresses the first research question:

RQ1: Does a data-driven approach to learning collocations facilitate EFL learners' development of collocational knowledge? If so, how does it facilitate such development?

Collocational knowledge was measured at three levels: the *receptive*, *controlled productive* and *free productive knowledge* of lexical collocations. The effects of DALC on developing collocational knowledge were evaluated through within-group and between-group comparisons at these three levels. Section 4.1 discusses the efficacy of DALC on the learners' *receptive* and *controlled productive knowledge* of lexical collocations. Section 4.2 details the quantitative and qualitative changes that occurred in the learners' *free productive knowledge* of lexical collocations, including the number, type, and acceptability of the collocations produced in writing assignments. While the sections above look into DALC from a summative point of view, Section 4.3 gives a formative account by examining the worksheets distributed in the DALC sessions. This chapter concludes with a summary of the ways in which DALC mediated the learning of lexical collocations (4.4).

### 4.1 Effects of DALC on the receptive and controlled productive knowledge of collocations

The *receptive knowledge* of lexical collocations (i.e., *recognition*) was operationalized as the score for MC questions on the pretest and post-test (see Appendices 7 and 8), and the *controlled productive knowledge* of collocations (i.e., *cued recall*) was measured by gap-filling questions on the same tests. Tests on lexical collocations were administered to the participants before and after the (non)intervention period, to measure the *changes* in their collocational knowledge. Of 109 participants in the experimental group, 101 participants took the pretest and 102 participants took the post-test, among which 97 participants overlapped by taking both the pretest and post-test. The difference in the number of participants taking tests was anticipated due to occasional absences of students. I, as a researcher, did not have control over the



attendance. For the purposes of intrapersonal comparisons, only the test scores of the overlapping 97 participants were submitted to statistical analysis. Similarly, the pretest was administered to 77 participants in the control group, and the post-test was given to 76 participants, so only the scores of the overlapping 76 participants who took both tests were included for analysis. Four sets of test scores (2 groups x 2 tests) were subjected to statistical analysis. Learner performances were compared within and between groups to understand the impact of DALC on receptive and controlled productive collocational knowledge.

### 4.1.1 Entry-level knowledge

Learners' *receptive* and *controlled productive knowledge* of lexical collocations were elicited by means of two collocation tests, a pretest and a post-test. This subsection discusses learner performances on the pretest. The table below shows the pretest scores:

*Table 4-1 Descriptive statistics for the pretest*

	Min.(%)	Max. (%)	Median (%)	Mean (%)	SD (%)
Exp. group (N=97)	18.00	60.00	34.00	35.30	8.82
Control group (N=76)	14.00	52.00	32.00	33.29	9.00

The experimental group scored 35.30% in the pretest, while the control group scored 33.29%. The pretest aimed to determine participants' entry levels of collocational knowledge, in order to provide information on the comparability of the two groups. The prerequisite for between-group comparisons is a comparable baseline: that is, the differences in the entry level between the two groups should be statistically insignificant, so as to provide an equal baseline for comparisons.

*Table 4-2 Between-group comparison of pretest scores*

	Median (%)	SD	U	Effect size (r)	Sig. (2-tailed)
Exp. (N=97)	34.00	8.82	3251.50	0.10	.183
Control (N=76)	32.00	9.00			

Note: Mann-Whitney *U* test (see Footnote 8, p.81).

As shown in the table above, the experimental group scored 34.00% (Mdn) in the pretest, and the control group scored 32.00% (Mdn), which was slightly lower than the median score of the experimental group, but did not reach a statistically significant level ( $p = .183$ ). This result provided a sound basis for between-group comparisons, because no statistically significant difference in the entry level was found between the

two groups. Note that there is one word in the pretest might be unknown to the participants, *dogged*. All the remaining words in the test are based on the 7,000 wordlist prescribed to high school students in Taiwan, and they are assumed to have mastered these words prior to their entrance to the university. The word *dogged* was chosen because it was listed in several collocation dictionaries as a typical adjective collocates of the noun *determination*, and the collocation *dogged determination* was thus selected to be one of the target collocations to be learned in the pedagogical mediation. Despite the likelihood of participants not knowing the word and the score of this particular test item being confounded, it was necessary to include the word in the pretest, so as to serve as a baseline of how well the participants know this target collocation, and to be compared with their knowledge of this target collocation after the intervention.

Before probing further into the effects of DALC on the learners' collocational knowledge, their entry-level knowledge merits discussion. In general, both groups performed unsatisfactorily in the pretest, scoring fewer than half of the test items ( $M=35.30\%$  and  $33.29\%$  respectively). As noted in Chapter 3 (3.5.2.1), the node words of the target collocations were selected from the prescribed wordlist in the participants' coursebook, so they were assumed to have mastered the node words by the time of the study, at least at the level of word meaning (i.e., form-meaning link). Given that the node words were studied previously, the result here reveals that the learners' knowledge of collocations lagged far behind that of word meanings. It seems that the number of lexical collocations the learners had at their disposal was considerably smaller than their vocabulary size. This finding is consistent with Bahns and Eldaw (1993): they found that their advanced EFL learners were more than twice as likely to produce miscollocations as they were to produce incorrect lexical words. They thus concluded that EFL learners' collocational knowledge fell behind that of general vocabulary. While Bahns and Eldaw (ibid.) only looked into VN lexical collocations, the present study expanded the scope of the investigation to encompass all types of lexical collocations, but found a similar gap between the learners' knowledge of collocations and that of word meanings. As noted in Chapter 2 (see 2.2.1 for word knowledge framework), word knowledge entails much more than word meanings, so a mere grasp of form-meaning links does not mean 'knowing' a word, and much less being able to use the word. Whereas it is relatively easier to improve

vocabulary knowledge, it is not an easy task to use the word productively (Nation, 2001). It can be argued that different aspects of word knowledge may not develop at the same rate (Laufer, 1997; Schmitt, 1998b), nonetheless, the huge gap found here between knowledge of word meanings and that of lexical collocations may cause considerable difficulties to the learner when it comes to putting the word into productive use. Such a gap may well be attributed to the misconception of learning words as discrete units and overlooking their syntagmatic relations. The pedagogical implications will be discussed in Chapter 7 (7.3).

### 4.1.2 Learner performance on collocation tests

The previous subsection has established that the two groups were closely matched in terms of the baseline performance on the collocation test. The comparison between pretest scores provided a sound basis for further comparisons between groups. In order to understand the effectiveness of DALC on the learners' collocational knowledge, the *changes* that occurred in the course of the (non)intervention period were examined. The table below shows learner performances on the pretest and post-test:

**Table 4-3 Descriptive statistics for the pretest and post-test**

		Min.(%)	Max. (%)	Median (%)	Mean (%)	SD (%)
Exp. (N=97)	Pretest	18.00	60.00	34.00	35.30	8.82
	Post-test	14.00	94.00	58.00	60.64	20.48
Control (N=76)	Pretest	14.00	52.00	32.00	33.29	9.00
	Post-test	10.00	72.00	36.00	36.24	11.21

The experimental group scored 35.30% in the pretest, and progressed considerably to 60.64% in the post-test. However, the standard deviation (henceforth SD) also increased with the test scores (the variation within the experimental group will be discussed later in this subsection). In contrast, the control group scored 33.29% in the pretest, and progressed only marginally to 36.24% in the post-test.

The tables below show within-group comparisons of test scores. The experimental group scored 34.00% (Mdn) in the pretest, and progressed to 58.00% (Mdn) in the post-test. The increase in test scores reached a statistically significant level ( $p < .001$ ). Notwithstanding the low entry-level performance, the experimental group seemed to benefit from DALC intervention, as evidenced by the substantial progress in the post-test.



**Table 4-4 Within-group comparison of test scores (experimental group)**

		Median (%)	SD	Z	r	Sig. (2-tailed)
Exp. group (N=97)	Pretest	34.00	8.82	-8.08	-0.58	.000*
	Post-test	58.00	20.48			

Note: Wilcoxon signed-rank test (see Footnote 9, p.81).

In contrast, the control group scored 33.29% (Mean) in the pretest and 36.24% (Mean) in the post-test (see Table 4-5 below). Despite a slight increase in test scores, the performances of the control group did not differ significantly over the non-intervention period ( $p=.082$ ).

**Table 4-5 Within-group comparison of test scores (control group)**

		Mean (%)	SD	T	DF	Sig. (2-tailed)
Control group (N=76)	Pretest	33.29	9.00	-1.77	75	.082
	Post-test	36.24	11.21			

Note: Paired sample t-test (see Footnote 10, p.81).

However, it is worth noting that the SD in the post-test scores of the experimental group had also widened with the increase in test scores. The large SD (20.48%) relative to the mean score (60.64%) seems to indicate highly varied performances across participants within this group. In other words, the experimental participants may not have benefited from DALC intervention to the same degree: some may have progressed considerably as a result of the intervention, while others much less so. Even though the overall trend was one that the experimental group as a whole performed considerably better on the post-test as a consequence of the pedagogical mediation, the high variability among participants was anticipated considering individual differences in learning. It is widely cited that learners respond differently to technology-enhanced language pedagogy due to differences in learning style or learner attribute (cf. Ford and Chen, 2000). Some learners may find that corpus consultation agrees with their learning preferences, while others are less enthusiastic about such an approach. For example, aligning the learners' test performances with their questionnaire responses (as will be discussed at length in Chapter 6), a negative correlation was found between the degree of progress in collocation tests and the response to Questionnaire Item 18 ( $r = -.217, p = .016$ ): that is, the less the learner found an inductive approach applicable to his/her vocabulary learning experience, the more progress that s/he made on the collocation tests. By an *inductive approach to learning vocabulary* I mean observing multiple instances in which the target word is embedded before generalizing how it is typically used in extended contexts, as

opposed to a *deductive approach* in which the usage of the target word is directly taught to the learner, so that s/he does not have to generalize. The result indicates that the learners who resorted to the inductive approach less frequently benefited more from DALC intervention. It may well be the novelty of encountering a wealth of language data for induction that made those who were less familiar with such an approach more sensitive to the intervention effects, hence a higher degree of progress. As plausible as such individual differences may seem, further investigation is needed to verify the extent to which they affect learner performances in a DALC setting. In addition, as noted earlier (see 4.1), occasional absences of students might also have led to varying degrees of intervention effects across participants.

In general, the two groups were statistically comparable before the intervention in terms of collocation recognition and cued recall: both groups scoring fewer than half of the items on the pretest. The experimental group benefited from DALC intervention, and made considerable progress afterwards. The control group's performance had not increased significantly over time. The respective changes in collocational knowledge over the intervention or non-intervention period attested to the positive effects of DALC on the receptive and controlled productive knowledge of lexical collocations. However, it is worth noting that the progress made by the experimental group may have been partly attributable to practice effects, because the test items were used as part of the intervention given to the experimental participants (see 3.5.2.3). Despite some modifications made on the post-test, the similar items used may have contributed to the experimental participants' familiarity with the items, and hence better performance in the post-test. While positive changes were found in the experimental group's collocational knowledge, it is necessary to note that such changes were not impervious to practice effects, due to the design of the intervention and data-elicitation instruments.

#### **4.1.3 Changes in the receptive and controlled productive knowledge of collocations**

As the control group did not receive DALC intervention, a further probe into the changes in the learners' collocation recognition and cued recall was thus limited to the experimental group. Receptive collocational knowledge (i.e., recognition) was

measured by MC questions, while controlled productive knowledge (i.e., cued recall) was measured by gap-filling questions. In terms of recognition, participants progressed from 43.43% to 68.18% (Mdn,  $p < .001$ ). With respect to cued recall, the median score increased from 25.93% to 57.14% ( $p < .001$ ). The results are summarized in the table below. DALC intervention seems to facilitate both recognition and cued recall of lexical collocations.

**Table 4-6** *Changes in the scores for MC questions and gap-filling questions*

		Median (%)	SD	Z	r	Sig. (2-tailed)
MC	Pretest	43.43	10.50	-7.97 <sup>a</sup>	- 0.57	.000*
	Post-test	68.18	17.36			
Gap-filling	Pretest	25.93	10.81	-7.72 <sup>a</sup>	- 0.55	.000*
	Post-test	57.14	25.74			

Note: Wilcoxon signed-rank test (see Footnote 11, p.82)

a. Based on negative ranks.

A closer look into the changes that occurred in recognition and cued recall shows that the latter increased to a greater extent than the former, as the median score for gap-filling questions doubled in the post-test. Throughout the intervention period, the participants' performance on recognizing, or distinguishing acceptable collocations from unacceptable ones, had been consistently better than that on recalling collocations cued by contextual information (MC 43.43% > gap-filling 25.93% in the pretest, and MC 68.18% > gap-filling 57.14% in the post-test). However, the latter in the post-test did not lag as far behind the former as it did before the intervention. In other words, the gap between these two levels of collocational knowledge reduced, because the degree of progress in cued recall was greater than that in recognition. It is likely that controlled productive knowledge was more sensitive to DALC than receptive knowledge: through corpus consultation learners had the opportunity to encounter the same target collocation in various contexts, so recalling collocations in cued contexts might be relatively easier. Alternatively, this might have been the result of *practice effects*, as the items measuring cued recall were almost identical in the two tests, with only one new item added to the post-test. In comparison, MC items might be relatively more difficult, since they were modified in the post-test: MC questions in the pretest consisted of two correct options and one incorrect option, while those in the post-test had one correct option and two incorrect ones. Although the items themselves remained the same, the options provided were different: it is likely that these modifications led to a difference in the level of difficulty and thus a smaller



extent of progress made in MC items, compared with that made in gap-filling items. Overall, it is reasonable to claim that DALC intervention had positive effects on the receptive and controlled productive knowledge of lexical collocations to a varying degree. Also, though the result shows that corpus consultation extended receptive and controlled productive collocational knowledge, it does not claim that such endeavours pushed receptive knowledge toward controlled productive knowledge along the receptive-productive continuum. The research design of this study does not allow such a claim, because the collocations tested in MC questions (receptive knowledge) were different from those tested in gap-filling questions (controlled productive knowledge) (see the design of test items in 3.5.2.1). Therefore, while DALC intervention contributes to receptive and controlled productive collocational knowledge respectively, whether it pushes collocational knowledge from the former to the latter remains uncertain from the current findings. This study has gone so far as to substantiate that DALC extends the above two levels of collocational knowledge respectively, it merits further research to investigate whether and how corpus consultation advances collocational knowledge along the receptive-productive continuum.

Chan and Liou (2005) compared the effects of web-based tuition with and without concordancing on EFL learners' knowledge of VN collocations (as discussed in 2.5.1.2). The results showed that web-based collocation instruction with concordancing was more effective than without concordancing, in terms of extending participants' VN collocational knowledge and raising collocational awareness. In Chan and Liou (ibid.), collocational knowledge was elicited by cloze questions, which measured only one aspect of collocational knowledge, namely *controlled productive knowledge*. More specifically, their results attested to the efficacy of concordancing only insofar as it extended the learners' *controlled productive knowledge* of VN collocations. The present finding coincides with Chan and Liou (ibid.) in that learners made significant progress in the *controlled productive knowledge* of lexical collocations after receiving DALC. Notably, the present study measured knowledge of all types of lexical collocation, rather than that of VN collocations alone. Taken together, corpus consultation is beneficial for developing the *controlled productive knowledge* of VN collocations as well as that of other lexical collocations. Interestingly, whereas the learners in this study worked with six designated web-based

concordancers, those in Chan and Liou (ibid.) were given access to only one of the concordancers, which was nonetheless sufficiently conducive to enhancing the learners' controlled production of VN collocations with adequate pedagogical mediation.

Likewise, Sun and Wang (2003) also looked into the efficacy of concordancing on collocation learning, with the exception that the collocations thereof were grammatical ones as opposed to lexical collocations (as reviewed in 2.5.1.2). The researchers found that *easy grammatical collocations* were more amenable for an inductive approach with concordancing, while *difficult* ones could be facilitated by either inductive or deductive approach. Taking together the present finding and that of Sun and Wang (ibid.), corpus consultation seems to facilitate the learning of *lexical collocations* as well as *grammatical ones*.

Corpus consultation also contributes to the learning of other aspects of word knowledge, in addition to that of collocations. Cobb (1999), for example, examined the efficacy of corpus consultation on the knowledge of word meanings (i.e., form-meaning links) through comparisons with that of wordlist/dictionary consultation (as discussed in 2.5.1.1). The results showed that the concordancing group outperformed the wordlist/dictionary group on the *recognition* and *cued recall* of the definitional knowledge of words. In particular, the concordancing group made significantly more progress in *cued recall*. In line with Cobb (ibid.), the present study also found that corpus consultation contributed significantly to both *recognition* and *cued recall* of collocations, and notably, the degree of progress in *cued recall* was higher than that in *recognition*. Corpus consultation is therefore beneficial for promoting the definitional knowledge of words as well as collocational knowledge. In general, it seems that corpus consultation is particularly conducive to enhancing the *controlled production* (i.e., cued recall) of words and/or collocations. It is likely that the observation and analysis of concordances (i.e., a query word occurring in multiple instances) is amenable to transferring *recognition* to *cued recall* in novel contexts. The profusion and variability of corpus data may enable the learner to go a step further from merely *recognizing* a word and/or collocation to being able to *recall* it in response to contextual cues. Cobb (ibid.) attributes this potential to the fact that corpora allow a target word to be encountered in *varied situations* in addition to

*varied contexts* (see 2.5.1.1 for elaboration). The observation of how a word is typically used in *varied contexts* under *varied situations* may lead to better controlled production of words and/or collocations in novel contexts.

Taking together the findings of the present and previous studies, corpus consultation or concordancing has been attested to facilitate the development of various aspects of word knowledge (e.g., word meanings/form-meaning links, grammatical and lexical collocations), particularly at the level of *controlled production*. In other words, corpus consultation has the potential to extend the learners' *definitional* and *collocational knowledge*, insofar as they are exposed to a predetermined context which requires a retrieval of appropriate word forms or collocations from their repertoire. However, previous studies have come only so far in investigating the efficacy of corpus consultation on the *controlled production* of collocations, and whether the effects extend to *free production* remains unresearched. To address the issues of free production of collocations, the section that follows will discuss how learners actually produced lexical collocations to express meanings of their own choice before and after DALC intervention.

## **4.2 Effects of DALC on the free productive knowledge of collocations**

Free productive collocational knowledge was measured by the lexical collocations used in the learners' writing assignments. As noted in Chapter 3 (see 3.5.2.2), the requirements (e.g., genre, topic and length) of the sampled writings varied according to the writing classes to which the learners had been assigned when they enrolled at the university. The genre of the assignments written by the experimental group was mainly *expository* with an average length of 200 words, while the control group was required to write *argumentative* essays of approximately 340 words (see 3.5.2.2 for the topics of the writings sampled). For the practical reasons stated in Chapter 3, I was only given partial access to this dataset, so the number of writing assignments collected was smaller than expected: 56 and 55 assignments were collected from the experimental group before and after the intervention, respectively; 44 and 46 assignments from the control group were sampled at the same points in time.

The lexical collocations produced in the writing assignments were manually identified,



based on a revised version of Benson, Benson and Ilson's (1997) classification of lexical collocations (see the working classification and the rationale for revision in 3.5.3.3). The lexical collocations identified were then classified according to their *types* along the collocational continuum (i.e., *free combination*, *restricted collocation* and *fixed expression*, see 3.5.3.4 for the classification). Furthermore, the *acceptability* of the lexical collocations was judged, and the factors contributing to *miscollocations* were considered (see 3.5.3.4 for judging acceptability and classifying miscollocations). Overall, the free production of lexical collocations was measured *quantitatively* and *qualitatively*, looking into the *number*, *type*, *acceptability* of the lexical collocations produced in learner writings. As the number and type of collocations used in writing may vary with genre, topic or length, it would be biased to make any *between-group comparisons* of collocations in this case. The following discussion on the changes in free productive collocational knowledge brought about by corpus consultation is thus limited to *within-group comparisons* only.

#### 4.2.1 The number of collocations

Based on the working classification of lexical collocations (see 3.5.3.3, Table 3-3), 1,200 lexical collocations were identified in the pre-treatment writings of the experimental group, and the number increased to 1,600 after the intervention. The control group used 1,838 lexical collocations in the writings prior to the non-intervention period, and the number increased to 2,169 afterwards (see below).

**Table 4-7 Collocation production in writing assignments**

		No. of writing assignments	Average length (words)	Total no. of collocations	No. of collocations per 100 words
Exp. group	Pre	56	198.23	1,200	11.10
	Post	55	202.64	1,600	14.57
Control group	Pre	44	325.02	1,838	12.99
	Post	46	367.63	2,169	12.77

As shown in the table above, the average lengths of the writing assignments varied, so it was more sensible to compare the number of collocations used per 100 words, rather than the raw number of collocations used in a piece of writing. On average the experimental group produced approximately 11.10 lexical collocations per 100 words before the intervention, and used 14.57 collocations per 100 words afterwards. As for the control group, the ratio of the number of collocations produced per 100 words

decreased slightly from 12.99 to 12.77 over the same period of time.

The table below shows that the experimental group made significant progress in the number of lexical collocations produced per 100 words ( $p < .001$ ), whereas the control group did not ( $p=.636$ ).

**Table 4-8 Within-group comparisons of the number of lexical collocations per 100 words**

		Mean	SD	T	DF	Sig. (2-tailed)
Exp. group	Pre	11.10	3.02	-6.35	52	.000*
	Post	14.57	2.96			
Control group	Pre	12.99	2.26	0.48	43	.636
	Post	12.77	1.84			

Note: Paired sample t-test (see Footnote 12, p.84).

The significant increase in the lexical collocations attempted by the experimental group may have been the result of being exposed to concordances and feeling more confident in experimenting with lexical collocations. Alternatively, it may well be attributed to the differences in writing topic. For the experimental group, though the overall genre of writing assignments was *expository*, the pre-treatment one focused specifically on *description*, while the post-treatment one targeted at *comparison* and *problem-solution* (see 3.5.2.2, Table 3-1 for the topics of writing assignments). Although these two sets of writing samples shared the overall *expository* genre, the differences in writing topic might have affected the use of lexical collocations to some extent, because collocates tended to cluster within certain semantic fields, as evidenced in recent corpus findings (Carter, 1998; Stubbs, 1995a). However, thus far, there is no empirical evidence to shed light on the association between genre types and the number of lexical collocations used. It is therefore unlikely that the extent to which writing foci influence the production of lexical collocations will be determined, be it the quantity or quality thereof. Hence, it is more reasonable to attribute the increase in the number of lexical collocations attempted partly to the learner's corpus consultation experience, while having in mind the influences of different writing foci. On the other hand, the ratios of collocations per 100 words remained stable over time on the part of the control group, which was likely to be the result of non-intervention and/or similar writing foci (i.e., arguing about a controversial issue, see 3.5.2.2, Table 3-1).

## 4.2.2 The types of collocations

To probe further into the learners' free production of lexical collocations, three types of word combination along the collocational continuum were distinguished, based on the degree of restrictedness between the components: *free combinations (FCs)*, *restricted collocations (RCs)* and *fixed expressions (FEs)* (see 3.5.3.4 for more methodological details of classifying lexical collocations; the RCs identified are listed in Appendix 15 and Appendix 16). Table 4-9 shows the frequencies and proportions of the three types of collocation produced by the learners:

**Table 4-9** *Descriptive statistics for the types of collocation used in writing assignments*

	Experimental group				Control group			
	Pre		Post		Pre		Post	
	No./100 words	Mean %	No./100 words	Mean %	No./100 words	Mean %	No./100 words	Mean %
FC	10.47	93.10	13.01	90.37	11.66	89.37	11.74	91.90
RC	0.62	6.86	1.54	9.45	1.27	10.09	1.01	7.92
FE	0.01	0.04	0.02	0.10	0.06	0.54	0.02	0.18
Total	11.10	≈100	14.57	≈100	12.99	≈100	12.77	≈100

The experimental group used 10.47 FCs per 100 words in the pre-treatment writings, and the ratio increased to 13.01 afterwards. Similarly, the ratio of RCs per 100 words also increased from 0.62 to 1.54 over the intervention period. As for FEs, there was only one instance found in the pre-treatment writings and two instances afterwards. The control group produced 11.66 FCs per 100 words before the non-intervention period, and the ratio increased marginally to 11.74 at the closure of this period. In contrast, the ratio of RCs per 100 words decreased from 1.27 to 1.01 over the same period of time. The ratio of FEs per 100 words also decreased from 0.06 to 0.02 over time.

Broadly speaking, the learners used a fairly high proportion of FCs, because such combinations were the least restricted on the collocational continuum. The constituent parts can be substituted freely as long as the resulting combinations are semantically acceptable and grammatically well-formed. Learners' heavy reliance on FCs found here lends support to the claim that non-native speakers construct a large proportion of language from rules instead of lexicalized routines, as pointed out by researchers such as Foster (2001) and Skehan (1998). To be precise, non-native speakers operate to a large extent on Sinclair's (1991) *open-choice principle*, which sees language as



constructed from grammatical rules into which discrete lexical items are to be slotted, as opposed to the *idiom principle*, which concerns the syntagmatic relations between words (see 2.1 for the two principles). The heavy use of FCs here was a manifestation of the *open-choice principle*. Such disproportionate collocation production seems to indicate that the learners may not have an adequate amount of RCs at their disposal, and thus need to construct utterances from scratch, as evidenced by their reliance on FCs. In the light of her empirical findings, Granger (1998) showed that French learners of English overused *general-purpose* or *all-round* amplifiers in adverb-adjective lexical collocations such as *completely different/new/free*, as opposed to more restricted ones (e.g., *bitterly cold*). In line with the present finding, her learners seemed to use amplifier adverbs as building bricks rather than parts of prefabricated patterns, which can also be understood as operating on the *open-choice principle*. She concluded that the learners' phraseological skills were severely limited, in that they underused native-like prefabs while overusing foreign-sounding ones. To compound the problem, Hill (2000) warns against the potential pitfalls that the more the learner creates utterances from scratch, the higher the likelihood that s/he will make collocational mistakes.

A possible explanation for the participants' heavy reliance on FCs was underdeveloped collocational knowledge. Given the participants' low entry level of receptive and controlled productive collocational knowledge (both groups scored lower than 40% as illustrated in 4.1.1), they had difficulty recognizing and recalling the frequent collocations of the words previously studied, much less putting them into productive use in writings. It is likely that learners had too limited a repertoire of RCs to use in written production, so they relied heavily on FCs. Moreover, it may well be that the requirements of writing assignments and the stakes involved had an impact on the use of collocations. The grading conventions in the writing classes may have favoured grammatical accuracy over lexical and/or collocational richness, thus implicitly discouraging taking risks to experiment with more restricted or error-prone word combinations. The writings were sampled from the participants' assignments of the writing class, the completion of which was not at all stake-free as they were part of the assessment in the course. As a consequence, instead of taking the risk to experiment with more restricted collocations, learners may prefer focusing their attention on getting meanings across in the most grammatically accurate way, so as to

reduce the likelihood of making mistakes. Therefore, they tended to rely on FCs rather than RCs, since the former was 'safer', entailing a lower risk of mis-collocating. As will be demonstrated in the next subsection (4.2.3), the acceptability of the collocations used in the writing assignments was fairly high, which may have been the result of such strategic use of risk-avoidance, rather than an indication of a developed sense of collocational knowledge.

As regards the degree of restrictedness of lexical collocations, Nesselhauf (2003) investigated 32 advanced EFL learners' use of VN lexical collocations in free written production, and found a similarly disproportionate profile of collocation use: approximately 79% of *free combinations*, 20% of *restricted collocations* and 1% of *idioms/fixed expressions*. The present study differed from Nesselhauf (ibid.) in that the participants in this study were less proficient EFL learners, and all lexical collocations in addition to VN ones were counted. Despite these differences, the similarly disproportionate profiles may be indicative of EFL learners' deficit of RCs in their repertoire. In similar vein, Howarth (1998b) compared the VN collocation use of native speakers of English in academic writing to that of 10 highly advanced EFL learners. He found that non-native speakers (NNS) used 69% of *free combinations*, 24% of *restricted collocations* and 1% of *idioms/fixed expressions*, compared with 63% of *free combinations*, 33% of *restricted collocations* and 5% of *idioms/fixed expressions* in the native speaker (NS) data. Although these two empirical studies differed from the present one in terms of the learners' proficiency level and the scope of lexical collocations investigated, the general trend was that the proportion of FCs decreased, while that of RCs increased as the learner became more proficient in the TL (the respective proportions of types of collocation found across studies are summarized in the table below).

**Table 4-10 Collocation production profiles across studies**

	Present study NNS production (pre-treatment, %)	Nesselhauf (2003) NNS production (%)	Howarth (1998b) NNS production (%)	Howarth (1998b) NS production (%)
FC	91.19	79	69	63
RC	8.58	20	24	33
FE	0.22	1	1	5
Total	≈100	≈100	≈94 <sup>a</sup>	≈100

a: miscollocations accounted for 6%.

The proportions of FCs and RCs underwent changes over the (non)intervention period

(see Table 4-11 and Table 4-12 overleaf). In terms of the proportions of the types of lexical collocation used in writings, as shown in Table 4-11, prior to the intervention 94.44% (Mdn) of the lexical collocations used by the experimental group belonged to FCs, but the proportion declined to 90.91% (Mdn) after receiving the intervention. The decrease in the proportion of FCs was significant at the level of .002. In contrast, the proportion of RCs had risen from 5.56% to 9.09% (Mdn) over the intervention period. The increase in the proportion of RCs was also statistically significant ( $p=.002$ ). Whereas the experimental group used fewer FCs and more RCs after the intervention, their control counterparts showed the opposite trend. They used a significantly larger proportion of FCs after the non-intervention period ( $p=.027$ ), increasing from 90.97% to 92.91% (Mdn). Accordingly, as shown in Table 4-12, the proportion of RCs had decreased from 10.09% to 7.82% (Mean), reaching a statistically significant level ( $p=.035$ ).

Although the profile of collocation use remained disproportionate throughout the intervention period, the experimental group nonetheless made more attempts to use RCs (Mdn 5.56%→9.09%) and produced fewer FCs (Mdn 94.44%→90.91%) in their writings after receiving the intervention. The changes in proportions may well be the result of a heightened awareness of collocational restrictedness, or collocability, fostered by the DALC experience. Since RCs were less flexible than FCs, the correct use thereof required a higher degree of awareness of collocability. It appears that the experimental participants were sensitized to the restrictedness of lexical collocations by the concordancing experience, and developed an awareness that certain words tended to collocate in a less free sense. The increased collocational awareness seemed to make the learners less reliant on FCs and more willing to experiment with RCs. In contrast, the reason why the control group used even more FCs (Mdn 90.97%→92.91%) and fewer RCs (Mean 10.09%→7.82%) in the writings after the non-intervention period was likely to be the result of using the risk-avoidance strategy noted earlier, or simply a lack of RCs at their disposal.



**Table 4-11 Within-group comparisons of the proportions of FCs and RCs in writing assignments**

			Median (%)	SD	Z	r	Sig. (2-tailed)
Exp. group	FC	Pre	94.44	6.38	-3.13	-0.30	.002*
		Post	90.91	6.74			
	RC	Pre	5.56	6.38	-3.07	- 0.30	.002*
		Post	9.09	6.77			
Control group	FC	Pre	90.97	6.22	- 2.22	-0.24	.027*
		Post	92.91	4.35			

Note: Wilcoxon signed-rank test (see Footnote 15, p.85).

**Table 4-12 Within-group comparison of the proportions of RCs in control group's writing assignments**

			Mean (%)	SD	T	DF	Sig. (2-tailed)
Control group	RC	Pre	10.09	6.11	2.17	43	.035*
		Post	7.82	4.22			

Note: Paired sample t-test (see Footnote 16, p.85).

### 4.2.3 The acceptability of collocations

In addition to the *types*, the lexical collocations used in learner writings were also classified based on their *acceptability*. By *acceptability* I took into account the typicality/commonness of a word combination as well as its acceptability for the context in which it was embedded. For example, *cute curtain* might not be a typical/common word combination, but was used in one of the participants' writing to describe the décor of her ideal room, so the combination could hardly be rated as *unacceptable*. Hence, in this study the acceptability of a collocation was judged on the basis of the typicality/commonness in language as well as acceptability in relation to the context in which it was embedded (see more methodological details of judging acceptability in 3.5.3.4). The miscollocations identified are provided in Appendix 17 and Appendix 18.

As shown in the table overleaf, the experimental group produced 1,200 lexical collocations before the intervention, 92 of which were miscollocations, reaching an acceptability rate of 94.44% (Mdn, see Table 4-13). The number of lexical collocations used increased to 1,600 after the intervention, but the number of miscollocations reduced to 80, attaining a rate of acceptability of 96.00% (Mdn). On the other hand, the control group produced 1,838 lexical collocations in the writings before the non-intervention period, 147 of which were miscollocations, so the acceptability rate was 93.14% (Mdn). The acceptability rate increased (Mdn 94.91%) as the control group made more attempts to produce lexical collocations (N= 2,169)

but erred less in number (N= 128).

*Table 4-13 Descriptive statistics for the miscollocations in writing assignments*

		No. of collocations	No. of miscollocations	No. of miscollocations per 100 words	Rate of acceptability (Median, %)
Exp. group	Pre	1,200	92	0.79	94.44
	Post	1,600	80	0.66	96.00
Control group	Pre	1,838	147	1.02	93.14
	Post	2,169	128	0.72	94.91

Both groups had fairly high degrees of acceptability as to lexical collocation production, because the vast majority of the lexical collocations in the writings were FCs that were not subject to collocational restrictions, as discussed in the previous subsection (4.2.2), hence a lower likelihood of making collocational mistakes.

As shown in Table 4-14, before the treatment 94.44% (Mdn) of the lexical collocations produced by the experimental group was acceptable, and the rate increased to 96.00% (Mdn) after the treatment. The difference in acceptability rate was significant at the level of .029. The results show that the learners did benefit from the pedagogical mediation with respect to the acceptability of lexical collocations produced in their writing assignments. On the other hand, the control group also made significant progress ( $p=.043$ ) over the same period of time: the rate of acceptability increased from 93.14% to 94.91% (Mdn). Taken at face value, the rates of collocational acceptability increased over time, with or without DALC intervention. However, a closer look into the types of collocation used may be able to account for the control group's increase in acceptability rate. As noted in the previous subsection (4.2.2), the control group used an even larger proportion of FCs (Mdn 90.97% → 92.91%) at the closure of the non-intervention period: it may well be that they used more FCs, and therefore had a lower likelihood of producing miscollocations.

*Table 4-14 Within-group comparisons of collocational acceptability rate*

		Median (%)	SD	Z	r	Sig. (2-tailed)
Exp. group	Pre	94.44	7.64	-2.19	- 0.21	.029*
	Post	96.00	4.09			
Control group	Pre	93.14	5.94	- 2.02	- 0.16	.043*
	Post	94.91	4.79			

Note: Wilcoxon signed-rank test (see Footnote 19, p.86).

As far as the experimental group was concerned, given the fact that the number of

lexical collocations per 100 words increased over the intervention (see 4.2.1), the likelihood of producing miscollocations should have increased accordingly. Contrary to this assumption, not only did the learners produce more lexical collocations after the intervention, but they also made fewer miscollocations. Such positive changes in the quality of collocation production were at least partly attributable to the corpus consultation experience. However, I would not go so far as to claim that such progress was a result of direct transfer from what participants learned from the intervention sessions, considering that no target collocations in the intervention were actually used in the writing assignments. Even though none of the lexical collocations presented in the intervention sessions were actually used in their written production, the fact that the proportion of RCs and the degree of collocational acceptability both increased significantly over the intervention showed that DALC did have an impact on the learners' free production of lexical collocations to some extent. Generally speaking, the increase in collocational acceptability seems to suggest that the corpus consultation experience led the learners to devote more attention to avoid miscollocations, hence a heightened awareness of collocability.

As noted earlier (2.5.1.1), most studies on the efficacy of corpus consultation only looked into the learner's receptive and/or controlled productive knowledge, few ventured into free production: Kaur and Hegelheimer (2005) was one of the few studies to examine the effects of concordancing on learners' free production of academic words. As with the present study, Kaur and Hegelheimer (ibid.) investigated the effects of concordancing at three levels of knowledge: *receptive*, *controlled productive* and *free productive knowledge*, except that their focus was the learning of *academic words*. They did not find any significant difference at the receptive and controlled productive levels between the concordancing and non-concordancing (dictionary consultation) groups. However, the concordancing group outperformed the dictionary group in free production of academic words, measured by the correctness of word use in a writing task. The researchers concluded that whereas dictionaries were as effective as corpora for consulting word meanings, corpora seemed more useful in assisting learners to put static knowledge of words into active use. The present finding was consistent with that of Kaur and Hegelheimer (ibid.) in that concordancing was beneficial in increasing the acceptability of lexical collocations used in written production. It may well be that observing how collocations were used



in multiple contexts (as presented in concordances) drew the learner's attention to the formulaic nature of language, thereby making them more cautious about their own production of lexical collocations.

#### 4.2.4 Miscolllocations

As noted in the previous subsection, the rate of collocational acceptability of the experimental group increased significantly after DALC intervention (*Mdn* 94.44%→96.00%,  $p=.029$ ), it is therefore worth probing further to see what *qualitative* changes took place over this period of time. The miscolllocations produced by the experimental participants and possible contributing factors thereof were identified to illuminate the *qualitative* changes in the learners' collocational knowledge, in addition to the *quantitative* ones noted above.

Two broad categories of miscolllocations were identified: those caused by *interlingual* factors and those caused by *intralingual* ones (see a list of miscolllocations in Appendix 17). As the names suggest, the former was attributed to L1 influences, while the latter was caused by the TL, such as synonyms, de-lexicalized words, analogous collocations, and so forth. The factors contributing to *interlingual miscolllocations* were *literal translation* and *deviant translation*, which will be discussed in Section 4.2.4.1. The causes for *intralingual miscolllocations* consisted of *using synonyms*, *analogy*, *paraphrase*, *using de-lexicalized words*, *synformy* and *grammatical error* (see 4.2.4.2). There were also two instances that could not be classified, and thus labelled *unintelligible combinations*. The numbers of various types of miscolllocation produced before and after the intervention are as follows:

**Table 4-15** *Descriptive statistics for the miscolllocations in the writing assignments of the experimental group*

		Pre-treatment writing		Post-treatment writing	
		No.	Percent (%)	No.	Percent (%)
Interlingual miscolllocation	Literal	21	22.83	10	12.50
	Deviant	16	17.39	0	0.00
Intralingual miscolllocation	Synonym	6	6.52	6	7.50
	Analogy	1	1.09	0	0.00
	Paraphrase	40	43.48	47	58.75
	De-lexicalized	3	3.26	12	15.00
	Synformy	3	3.26	1	1.25
	Grammatical	0	0.00	4	5.00
Unintelligible		2	2.17	0	0.00
Total		92	100	80	100

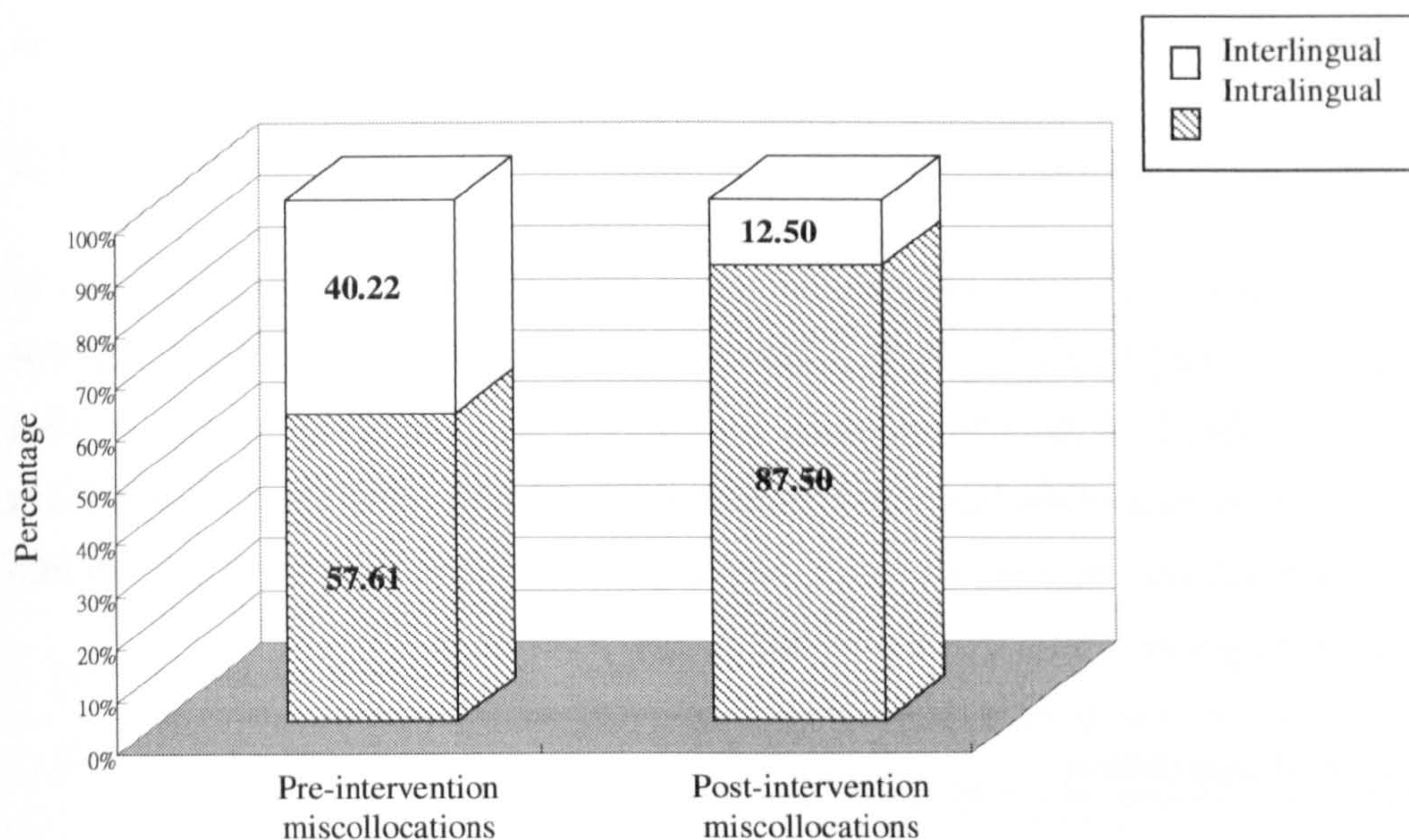


In terms of *interlingual miscollocations*, those caused by *literal translation* accounted for 22.83% of the miscollocations produced in the pre-treatment writings, and the proportion decreased to 12.50% in the post-treatment production. The proportion of the miscollocations resulting from *deviant translation* plummeted from 17.39% to none over the intervention.

As regards the factors contributing to *intralingual miscollocations*, *using synonyms* accounted for 6.52% of the miscollocations produced in the pre-treatment writings and 7.50% of those produced afterwards. There was one example of drawing on *analogy* in the pre-treatment writings, but no instance was found afterwards. A considerable proportion (43.48%) of miscollocations used prior to the intervention was due to *paraphrase*, and the proportion increased to 58.75% after the intervention. The proportion of *using de-lexicalized words* increased from 3.26% to 15.00% over the intervention period. Miscollocations caused by *synformy* decreased from 3.26% to 1.25% over the same period of time. Lastly, there were four miscollocations in the post-treatment writings that were caused by grammatical errors, accounting for 5.00% of the miscollocations produced in the post-treatment writings.

Overall, as shown in Figure 4-1, *interlingual miscollocations* accounted for 40.22% of the miscollocations produced before the intervention, but decreased to 12.50% afterwards. On the other hand, miscollocations resulting from *intralingual* factors increased from 57.61% to 87.50% over the same period of time. As noted in the previous subsection (4.2.3), the acceptability of lexical collocations produced in learner writings increased over the intervention period (*Mdn* 94.44%→96.00%): that is, learners made significantly fewer miscollocations after the intervention (0.79→0.66 miscollocations per 100 words).





**Figure 4-1** The miscollocation profile of the experimental group

More specifically, a closer look into the miscollocations produced in the two sets of writings reveals that not only did the number of miscollocations reduce, but the types of miscollocation produced also varied. The most marked difference between the miscollocations produced before and after the intervention was the changes in the proportions of *interlingual* and *intralingual miscollocations*. It seems that the learners relied to a greater extent on L1 resources for producing collocations before the intervention, but they were more inclined to use *intralingual* resources after having the corpus exploration experience. DALC intervention may have drawn the learners' attention to the *language-specific nature* of collocations (see 2.2.2.2), so they became more alert regarding the free interchangeability of word combinations on the sole basis of L1 translation. Even though the attempts to use *intralingual* resources for producing collocations were still less than successful, it was nonetheless a sign of progress, because the learners were less reliant on L1 and more willing to experiment with the TL, suggesting an increased awareness of the arbitrariness of collocation. Admittedly, DALC intervention did not 'miraculously' prevent the learners from producing miscollocations, nevertheless, it seemed to make them more cautious about the potential pitfalls of cross-linguistic transfer. In other words, the corpus exploration experience may have heightened the learners' awareness of restricted collocability. The following subsections illustrate the types of miscollocation identified in the



current dataset.

#### 4.2.4.1 Interlingual factors

L1 influence has been reported to account for a considerable proportion of miscollocations produced by EFL learners (Bahns, 1993; Gabrys-Biskup, 1992). The present finding is in line with the body of empirical evidence in this respect. It is not uncommon for learners to resort to L1 when they did not have the target collocation at their disposal nor other ways around. *Interlingual miscollocations* identified fall into two subcategories: *literal translation* and *deviant translation*.

##### 1) Literal translation

Approximately 1/5 of the miscollocations found in the pre-intervention writings could be attributed to *literal translation* from L1. The miscollocations in the following instances were caused by literal, word-for-word translation from the learners' L1, Mandarin Chinese.

(1) *\*sunshine sprays*

*Last, the inmost of my room will be a French Window. Imaging the lovely sunshine sprays from the window every morning. How peaceful!*

(2) *\*sunshine sprinkles*

*I have been dreaming of owning a perfect room located at beach. When I enter the room, sunshine will sprinkle on me.*

(3) *\*obey one's idea*

*People always do what they like do, and avoid unpleasant things. You should know everything won't always obey you idea.*

The miscollocations in examples (1) and (2) resulted from the literal translation from Mandarin Chinese, 陽光灑落. It seems that *sunshine* was conceptualized as a liquid in Mandarin Chinese, so that it could be *sprayed* or *sprinkled*. Interestingly, the same metaphor was used in English, but with a different range of verb collocates, such as *sunshine floods/pours/streams*. Although the two languages shared a conceptually parallel metaphor, the manifestations were nonetheless different, and likely to cause difficulties or confusion for learners. The intended meaning in (3) was *things don't always go as you wish*. The phrase *as you wish* should be the closest translation of the miscollocation *\*obey one's idea*, which was literally transferred from Mandarin Chinese, 順某人的意. The number of miscollocations caused by literal translation reduced by half over the intervention, indicating a heightened awareness of the

language-specific nature of collocation.

## 2) Deviant translation

In addition to literal translation, there were instances also caused by cross-linguistic transfer, but the transfer seemed to be confounded by an injudicious use of bilingual reference tools. In such instances, one lexical item in Chinese may encompass several distinct words in English, so the learner may have mistaken them as interchangeable, and thus produced miscollocations.

(4) *\*indiscriminate bathroom*

*I go through her indiscriminate bathroom; I see the garbage is full of the garbage can.*

(5) *\*transpire a smell*

*It is a bad habit because she gets unwashed clothes and takes it on; maybe it will transpire bad smell.*

In example (4), while *indiscriminate* and *messy* share the same Chinese translation 雜亂, they were neither synonymous nor interchangeable in English as a modifier. Moreover, using the less common adjective *indiscriminate* instead of the more common and contextually-appropriate ones such as *messy* or *untidy* was likely to be the result of an injudicious use of bilingual reference tools. The learner may have consulted bilingual dictionaries or thesauri for the English equivalents of the Chinese lexical item 雜亂, and may have been given a number of English words from which the learner chose a seemingly 'big word' without knowing the precise word meaning and range of frequent collocates.

As for example (5), *transpire* denotes the scientific phenomenon by which the water in a plant passes from its surface into the air. However, the word shares the Chinese translation 散發 with other verbs such as *emit* or *give (off)*. As a result, the learner was likely to find the word through bilingual reference tools, and used it regardless of the semantic and collocational restrictions.

The problematic choice of collocates above could be attributed to learners' partial lexical knowledge or injudicious use of reference tools. If the learner already had the erroneous collocates in his/her repertoire but misused them in the instances above, it was likely that s/he had only limited knowledge of these words. S/he might know the approximate meanings of words, but was unaware of the precise meaning senses and

range of typical collocates. In this case, the implication for pedagogy is that vocabulary needs to be taught or learned in conjunction with 1) frequent collocates to highlight the nuances between a word and its near-synonyms, because different ranges of collocates can be self-explanatory of the nuances in meaning senses; and 2) the extended contexts in which the word is embedded, as they may provide more semantic and stylistic information than teaching and/or learning the word in isolation. Alternatively, the reason why learners came up with such miscollocations might be an injudicious use of reference tools. Some learners might have accepted indiscriminately whatever was given by these tools without careful examination of the grammaticality or appropriateness, and thus produced problematic word combinations, for example:

(6) *\*form the falling head*

*There must be the first and the last in the same class because each student has different comprehension and studying ability. Besides, they will form the falling head between themselves. For those students who have bad comprehension, they are hardly to achieve too much homework.*

An unusual, yet grammatically well-formed word combination *\*form the falling head* was identified in the text above. Given the context, the combination *\*form the falling head* seems to express the meaning of 'there is a gap between the high- and low-achievers'. However, neither *\*form the falling head* nor *falling head* were found in a number of major English dictionaries. I then consulted an online bilingual dictionary, which was observed (in the think-aloud sessions) to be a reference tool on which the learners relied heavily. The search yielded a result in Chinese 形成落差, literally meaning *form a gap*. Interestingly, *falling head* was a specialized term used in the field of civil engineering, sharing the same lexical item 落差 with *gap* when translated into Chinese. Since none of the major dictionaries included such a highly domain-specific term, and the learner was an English major instead of a civil engineering one, the unusual word combination was very likely to be the result of an injudicious use of bilingual reference resources.

As will be demonstrated in Chapter 6 (6.2), a considerably high proportion of participants used online dictionaries and electronic ones (91.80% and 80.60% respectively) as reference tools to aid English learning. Only half of the respondents used conventional printed dictionaries. Specifically, observations in the think-aloud



sessions show that participants relied heavily and almost exclusively on two particular online bilingual dictionaries, Google and Yahoo dictionaries. While they may be handy tools to access word meaning, they nonetheless fall short of providing comprehensive and elaborate information on the nuances in meaning senses and common collocations. To compound the problem, credulous learners tend to rely heavily on the translation provided by these reference tools, which have been infamous for word-for-word, out-of-context machine translation, disregarding factors such as collocability, immediate contexts, discourse or genre/register. As illustrated in the example above, the learner might have queried in Chinese for an English equivalent on less authoritative reference tools, and accepted the quick and easy answer without further verifications. Notwithstanding the grammaticality of the word combination and its possibly acceptable collocability within certain specific fields, it failed to convey the intended meaning of the writer and might cause confusion for readers. The instances discussed above have important implications for FL pedagogy to introduce learners to adequate resources for accessing collocational information, to guide them to make judicious use of various reference tools, and to raise their awareness of the strengths as well as the limitations of such tools.

In sum, there were 16 instances caused by *deviant translation* in the pre-treatment writings, but no instances were found after the intervention. It may well be that the participants became more aware of the *language-specific* nature of collocation and the arbitrariness in the choice of collocates insofar as they had reservations about accepting whatever was given by the reference sources as true without further verifications. This is not to claim that reference tools other than concordancers are inadequate: they serve different purposes from concordancers, as they may provide information on word meaning or synonym, while concordancers are more suitable for consulting collocational information. As will be demonstrated in Chapter 6 (6.5.2.1), the participants also found concordancers to be better tools for obtaining information on *word use* (e.g., grammatical function and collocation), as opposed to that on *word meaning*.

#### 4.2.4.2 Intralingual factors

*Intralingual miscollocations* could be attributed to the following factors: *using synonyms, analogy, paraphrase, using de-lexicalized words, synformy and grammatical errors*. In such instances, the sources to which learners resorted lay in the TL, as opposed to L1 discussed in the previous subsection.

##### 1) Using synonyms

Near-synonym is cited to be one of the most common contributing factors of miscollocations (Farghal and Obiedat, 1995; Zughoul and Abdul-Fattah, 2003). For some near-synonyms, the nuances in meaning sense may not be entirely clear from the prescribed definitions, but need to be differentiated through an observation of how they are typically used with other words. As suggested by Gabrys-Biskup (1992), the more synonyms a word has, the more difficult it is for learners to distinguish between the synonyms and produce commonly used collocations. The instances below demonstrate the erroneous use of synonyms:

(7) *\*blond door*

*Then, there are three special door, including red door, green door, and, blond door.*

(8) *\*pay time*

*He likes to pay some times on picking up books and reading it as a time-killing thing, and so do I.*

Whereas *blond* and *golden* are near-synonyms as they may denote the same colour, they by no means modify interchangeably. This is a perfect example showing that the nuances of synonyms can only be told from *word use* rather than *word meaning* alone. The learner may have partially acquired the definitional knowledge of the word *blond* without knowing its range of collocates, which was also part of word knowledge but could only be made clear by the syntagmatic relations of the word. In (8) the verb *pay* was mistaken as interchangeable with *spend*, as they may be semantically similar in the sense of using money. However, as far as the referent *time* was concerned, *pay* and *spend* were by no means synonymous. As Farghal & Obiedat (1995) point out, the use of synonyms is evident of direct application of Sinclair's (1991) *open-choice principle*, since collocations may not be seen as chunks but as composed of slots to be filled in with lexical items. This type of miscollocation seemed to be caused by an overgeneralization of the collocability of synonyms.

## 2) Analogy

Some miscollocations are attributed to analogy, for example, the combination *\*pay notice* may be an analogy from the common collocation *pay attention*. The nodes here are falsely assumed to share the verb collocate *pay* and hence the miscollocation.

### (9) *\*take hot spring*

*In the lavatory, there is good views to see whole city and I can take hot spring if I don't want see these views.*

Likewise, in (9) the miscollocation *\*take hot spring* may have been an analogy from *take a bath*. Since the two nodes *spring* and *bath* share the sense of immersing one's body in hot water, it was likely that the learner mistakenly assumed that they also shared verb collocates. However, it was also possible that the learner used the verb *take* because it was a de-lexicalized word.

## 3) Paraphrase

Where no L1 or L2 equivalents can serve as a point of reference to express the intended meaning, learners may attempt to paraphrase their way around in order to achieve the goal of communication.

### (10) *\*resist fear*

*They need to resist their fear, and they should try to do tasks rather than quit.*

The intended collocation in (10) is most likely to be *overcome fear*. The miscollocation did not result from L1 transfer, and the verbs *overcome* and *resist* were semantically distinct and barely synonymous. The miscollocation *\*resist fear* may have been an attempt of paraphrasing to convey the meaning of dispelling fear and not to be defeated. As Lewis (2000) suggests, learners may encounter 'intermediate plateau' where they can paraphrase around to achieve the aim of communication, but they would not make much progress thereafter unless they extend collocational knowledge. Bahns & Eldaw (1993), on the other hand, see no harm in paraphrasing for communication purposes, and suggest that only collocations that cannot be paraphrased or circumvented merit pedagogical attention.



#### 4) Using de-lexicalized words

De-lexicalized words are what Dörnyei (1995) refers to as 'all-purpose words', which are taken to have little or no meaning in their own right, but depend semantically on the nodes with which they collocate (e.g., *thing*, *stuff*, *make* or *do*). Using de-lexicalized words was one of the strategies to which learners resorted when specific collocates were lacking.

(11) *\*make damage*

*In fact, I even can't believe that a earthquake could make this horrible damage.*

In the example above, the learner used the verb *make* to collocate with the noun *damage*, since s/he did not have the common collocation *cause/do damage* in the repertoire, and the de-lexicalized word *make* seemed to be a convenient solution.

#### 5) Synformy

Synformy, or similarity in lexical forms, may confuse learners and lead to miscollocations if they do not have solid knowledge of the target collocations. The learner may know a collocation to some extent, but is unable to produce the collocate in its precise form due to the confusion caused by orthographical similarity.

(12) *\*adopt in the environment*

*However, I can adopt in the environment quickly.*

In (12), the verb *adopt* may have been mistaken for *adapt* and hence the miscollocation *\*adopt in the environment*.

#### 6) Grammatical errors

There were four instances of miscollocation caused by confusion over the transitivity of the verb *procrastinate*, hence the miscollocations *\*procrastinate the thing/work*.

To sum up, the proportion of *intralingual miscollocations* increased considerably over the intervention period (57.61%→87.50%), while that of the *interlingual* ones decreased (40.22%→12.50%). It seems that DALC not only increased the proportion of RCs (see 4.2.2) and the acceptability of collocations (see 4.2.3), but also reduced the proportion of *interlingual miscollocations*. It may well be that DALC intervention sensitized the learners to the linguistic differences between the two languages, and

heightened the awareness that collocations were so *language-specific* that cross-linguistic transfer may not always be successful. Given the fact that the number of miscollocations, notably interlingual ones, reduced after the intervention, it is reasonable to attribute the positive impact to the corpus consultation experience, which directed the learners' attention to the *usage-based* and *language-specific* nature of collocations.

Overall, Sections 4.1 and 4.2 have discussed the effects of DALC across groups: the experimental group made considerable progress on receptive, controlled and free productive knowledge of lexical collocations, whereas the control group did not progress to a statistically significant extent. However, it is noteworthy that the contrast may have partly resulted from different amounts of time spent on each treatment. The experimental group was given approximately 30 minutes per session to work on the DALC worksheet, while the target collocations taught to the control group were embedded in vocabulary teaching, the time of which was not specifically controlled in order to remain as naturalistic as possible. It is likely that the more intensive exposure to the target collocations led to better performance on the part of the experimental group. Therefore, the difference in performance between the two groups may have been partly attributable to the different amounts of time the participants spent on the target collocations.

### 4.3 Formative evaluation

In addition to the summative evaluation of learners' collocational knowledge outlined above, this study also looked formatively into how they performed in the five weekly DALC sessions. The experimental group was given collocation learning worksheets (see Appendix 9) to work with web-based concordancers in these sessions. Five worksheets were distributed and then collected at the closure of the sessions to examine the processes in which the participants exploited corpus resources to learn collocations (see details of DALC intervention in 3.4). As shown in the table below, the experimental participants scored 92.71% on Worksheet 1, 50.53% on Worksheet 2, 83.04% on Worksheet 3, 94.55% on Worksheet 4 and 83.00% on Worksheet 5.

**Table 4-16** *Descriptive statistics for collocation learning worksheets*

Worksheet	No.	Min.(%)	Max.(%)	Mean (%)	SD
1	96	40.00	100.00	92.71	10.90
2	82	6.25	100.00	50.53	31.37
3	85	50.00	100.00	83.04	12.63
4	78	66.67	100.00	94.55	7.58
5	80	50.00	100.00	83.00	10.36

Apart from Worksheet 2, learner performance was satisfactory, considering the high scores obtained in general (over 80%). With the help of the designated web-based concordancers, learners were generally able to find target collocations. Nonetheless, the mean score of Worksheet 2 (50.53%) stood in stark contrast to other worksheets. Also, the SD of Worksheet 2 was fairly high (31.37% relative to the mean score 50.53%), suggesting a considerable variation in learner performance within the group. Since the web-based concordancers as a collocation reference tool worked effectively in all the other sessions, as evidenced by the mean scores, the slump in the score for Worksheet 2 needs a further probe. It is likely that the gap-filling questions in Worksheet 2 contained more gaps (i.e., 17 gaps) to be filled in, compared with other worksheets (i.e., 10-12 gaps). The number of collocations for concordancing may have caused more difficulty for the participants. Another possible explanation is the temporary breakdown of one of the web-based concordancers, *Collins WordbanksOnline*, at the time when half of the experimental participants carried out the task. The concordancer worked normally when the first half of the participants undertook this task, but broke down as the second half tried to access it. Apart from this, the participants did not experience any difficulty accessing the other five designated concordancers. To put it into perspective, the participants who did not experience the technical difficulty in accessing *Collins WordbanksOnline* scored 75.42%, while their counterparts scored merely 29.00% on the same worksheet. It seems clear that the temporary breakdown of this particular concordancer did have an evident impact on learner performance on Worksheet 2. On the other hand, the first half of the participants who had access to all the web-based concordancers did not score as high as they did on the other four worksheets, suggesting that the increase in the number of collocations for concordancing did have an impact on the completion of the worksheet, albeit not as considerable as that created by the breakdown of *Collins WordbanksOnline*.

Although six web-based concordancers were introduced to the participants, they



tended to rely heavily on *Collins WordbanksOnline*, as will be demonstrated in Chapter 6 (6.5.3.2): that is, 66.33% of the respondents found this tool helpful. Such a high degree of preference for this particular concordancer was closely related to the poor performance on Worksheet 2 as a result of the temporary breakdown. The reason for participants' favourable attitude toward this particular concordancer might be that it provided an advanced search function, which allowed the user to specify the word class of the query word and/or its collocates. This search function is efficient in narrowing the search scope, compared with that of simply querying the node word and trying to locate a particular collocate among a wide range of collocates (corpus query methods will be discussed in 5.1.1).

In sum, some participants' unsatisfactory performance on Worksheet 2 could have been the result of an unmanageable number of items on the worksheet, compounded by limited support from the reference tool on which they relied heavily. Broadly speaking, the participants' DALC experience was fairly successful, as they were generally able to locate target collocations with the assistance of web-based concordancers.

#### 4.4 Summary

This chapter has demonstrated the *learning product*, or *learning outcomes*, of DALC, as evidenced by the *quantitative* and *qualitative* changes that occurred in the learners' *receptive*, *controlled productive* and *free productive knowledge* of lexical collocations. With adequate concordancing skills, web-based corpora can be useful resources for providing information on *word use*, such as collocations. The findings suggest that DALC has a positive impact on the *receptive* and *controlled productive knowledge* of collocations: that is, the learners recognized and recalled significantly more lexical collocations as a result of self-access corpus exploration. Notwithstanding an overall progress in the above two respects, a considerable variation among learners warrants further investigations into the relationship between individual differences and the extent to which a learner benefits from such a pedagogical approach. As regards the *free production* of lexical collocations in writing assignments, the DALC experience increased the *number* and *acceptability* of lexical collocations. Notably, the experimental group produced a smaller proportion of FCs and a larger one of RCs

after the intervention, suggesting a heightened awareness of collocational restrictedness, or collocability. Moreover, they seemed to be sensitized to the language-specific nature of collocations, as they relied less on cross-linguistic resources in producing collocations after the intervention. Hence, DALC is conducive to 1) developing all levels of collocational knowledge; and 2) raising collocational awareness. The next chapter will probe further into the processes with which learners engaged during the DALC task.

## Chapter 5 Exploring Learning Processes of DALC

The previous chapter has demonstrated the learning outcomes of DALC, and this chapter proceeds to the learning mechanisms underlying DALC by providing a descriptive account of how learners fared cognitively with language data contained in corpora as a means of inducing target collocations. It addresses the second research question:

RQ2. What is the nature of the thinking processes EFL learners engage in during the data-driven approach to learning collocations?

Section 5.1 outlines the corpus consultation behaviour elicited from the participants' corpus query records. Section 5.2 and Section 5.3 detail learners' use of cognitive and metacognitive strategies to approach the DALC task. Instances of how learners deployed these strategies to process corpus data in order to induce collocations will be described at length. This chapter concludes with a summary of how learners coordinated mental capacities and mediating tools to approach the DALC task (5.4).

### 5.1 Corpus consultation behaviour

There were 17 participants in the experimental group volunteered to take part in the think-aloud sessions after the intervention period (see 3.6.1.4 for collecting think-aloud data in the main study). The participants were asked to answer the collocation questions on a worksheet (see Appendix 10) with the assistance of web-based concordancers or any resources available online. Meanwhile, the respondents were asked to verbalize their thoughts simultaneously as they undertook the collocation task. Participants' concurrent verbal reports (i.e., mentalistic data) were audio-recorded, and their queries on reference tools (i.e., behaviouristic data) were manually documented by myself.

The use of corpus resources to induce collocations makes the DALC task a *resourcing* endeavour by nature. *Resourcing* is one of the cognitive strategies identified by O'Malley and Chamot (1990:138): '[u]sing available reference sources of information about the target language, including dictionaries, textbook books, and prior work'. In the present study, the DALC task is in itself a *resourcing* activity, however, the information provided by corpora is not as straightforward and organized as that by



dictionaries or thesauri, so the user may need to manipulate corpus data physically or cognitively in order to obtain the information needed. It was therefore assumed that in the DALC task, the learner would employ mental capacities to process language input found in web-based resources as a means of inducing target collocations. The following subsections give an overview of the observable referencing behaviour in the DALC task, including the query methods employed and the reference resources consulted.

### 5.1.1 Corpus query methods

The six designated web-based concordancers are as follows:

Collins Wordbanks *Online* English Corpus Sampler  
IWiLL Collocation Explorer  
Lexical Tutor Corpus Concordance English  
NTNU Web Concordancer  
TOTALrecall  
VLC Web Concordancer

The concordancers vary in corpus size, genre/register and search functions offered (see Appendix 5 for a detailed description of the concordancers). Also, in the think-aloud sessions, the respondents were allowed to use any resources available online, in addition to the concordancers above, in the hope that the consultation endeavours would resemble the real-life ones as closely as possible. (the additional online resources the participants consulted and the ways in which they were used will be discussed in 5.1.2). When consulting corpora, the respondents generally employed the following query methods:

- 1) **Node search** is querying the node word of a target collocation: for example, querying the noun [economy] to search for its verb or adjective collocates in the immediate contexts.
- 2) **Collocate search**, on the other hand, is querying a candidate collocate that the respondent has in mind to see if it goes with the node word: for instance, querying the adjective [booming] to see if the node word *economy* falls within the range of the nouns it modifies.

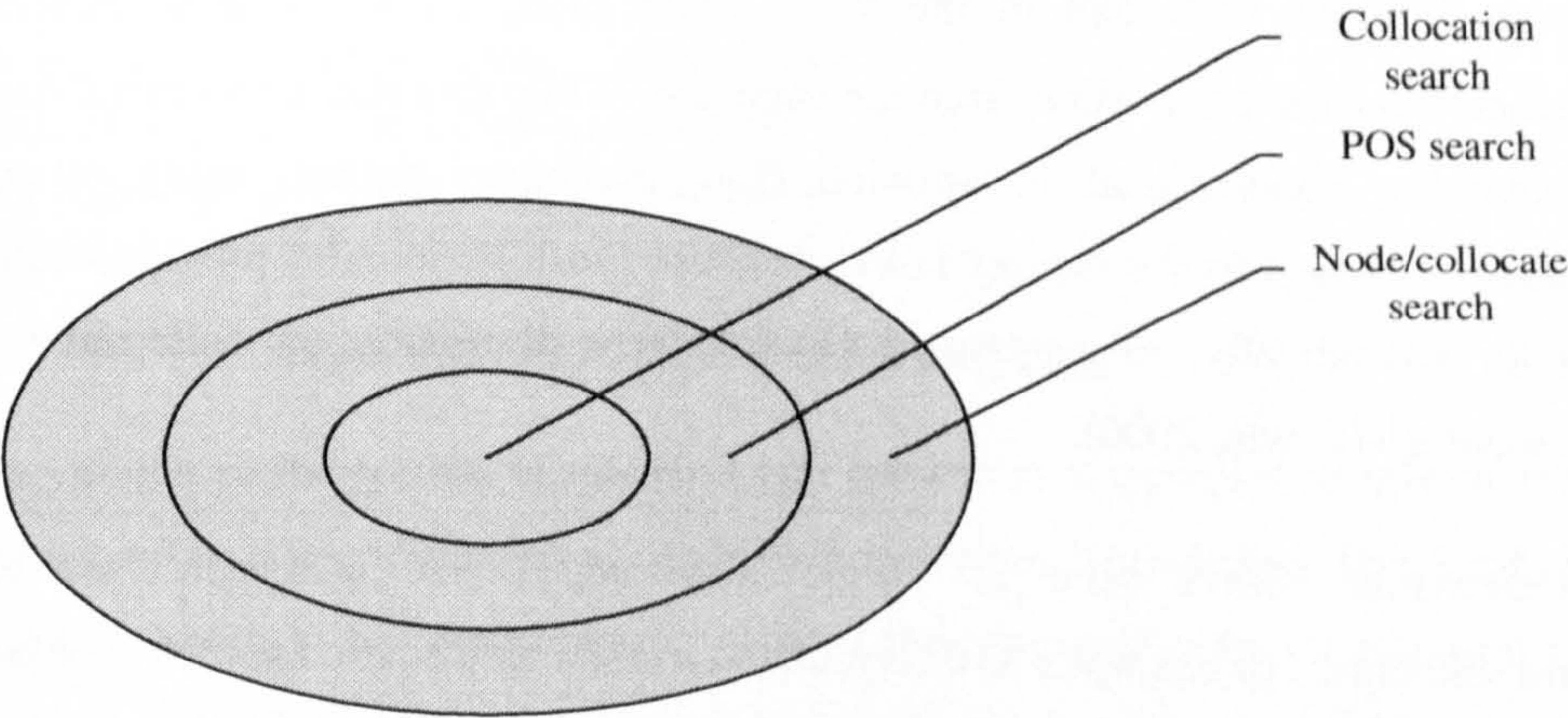
The above methods that query only a single component of a collocation usually



generate a substantial number of concordance results, whereas the following methods narrow down the search scope by specifying the word class of query words or by querying the exact candidate collocations in mind.

3) **POS (part-of-speech) search** specifies the word class of the query word(s) to focus the scope of search: for example, the query [JJ+economy] on Collins WordbanksOnline (JJ is the POS tag for adjectives in this concordancer) limits the range of collocates to adjectives only. By reducing the search scope to adjective-noun collocations, the learner is able to focus attention on the adjectives frequently modifying the noun *economy* without being distracted by collocates that belong to other word classes.

4) **Collocation search**: in cases where learners already have candidate collocations in mind, they may query the exact candidate collocations, such as [booming economy], to determine its collocability by checking this particular combination against real language data in corpora.



*Figure 5-1 Query methods and corresponding search scope*

Instead of relying on any single query method throughout the task, the respondents were flexible in alternating among query methods, as the situation demanded. The number of search results decreases with the refinement of queries. In other words, the more specific a query is, the narrower the search scope becomes (see visual representation in Figure 5-1).



5.1.2 Additional reference tools

In addition to the designated web-based concordancers, the participants were also allowed to make use of any online reference tools available to them. Further reference tools found to be employed by the respondents are as follows:

Google (search engine)	Yahoo Dictionary
Google Translate	Yahoo Knowledge
Yahoo (search engine)	Yahoo Translate

The respondents were found to consult online dictionaries for the definitions of unknown words in the worksheet or in concordances. There were also instances in which the respondents searched the sample sentences in online dictionaries for target collocations, for example, Student 4 (henceforth S4) found the target collocation *growing economy* embedded in a sample sentence in the entry *economy*:

The nation's **economy** is **growing** rapidly.

Although the syntactic relation between *growing* and *economy* in the sample sentence was different from that of the target collocation, S4 was able to generalize the adjective-noun collocation from the sentence. While this was a successful example of obtaining collocational information from dictionary entries, many other similar attempts found in the dataset failed, because most dictionaries provided collocations only sporadically, as contended by scholarly discussion on collocation learning resources (Lewis, 2000).

Discussion boards such as *Yahoo Knowledge* were also consulted by some respondents. For example, S10 formulated a query in Chinese [使用方法的英文] (use a/the method in English), however, the Chinese translation of the collocation *use a/the method* is the same as that of the word *instruction* (for using a product):

*As for Yahoo Knowledge, I'd enter this (collocation) first, and then I'd add a few words saying 'English for these words'. Then, people on the Internet, they will first provide some information individually, and I would examine (the results) one by one. Of course I won't be able to read through all of them because there is little time. [...] I see a lot of... keywords used for introducing products. This may not be what I wanted, so I won't keep looking.*

It is worth noting that this tool is a Chinese-medium one, so the majority of search results is not generated from authentic texts where English is used for real-life communication purposes, rather, they mainly come from online discussion on




Chinese/English translation. For instance, S16 queried [穩定的收入\_英文] (stable/steady income \_ English) and found one result showing:


知識+ 首頁 > 教育學習 > 語言文字 > 英文

用手機上知識+


分享 收藏至知識團 加入追蹤 轉寄朋友 友善列印


**知識問題** **收入不穩定的英文?**

 發問者：考試一百分 (初學者 5 級) [檢舉]  
發問時間：2006-04-19 01:41:05  
解決時間：2006-04-20 01:09:30  
解答贈點：5 (共有 0 人贊助)  
回答：3 評論：0 意見：0

網友正面評價  
**73%**   
(共有 11 人評價)

請問英文達人收入不穩定 有業績壓力 競爭對手多 有時候會遇到一些不禮貌的客人的英文怎麼講呢?

 **最佳解答** 發問者自選

 回答者：chieh98\_w (實習生 5 級) [檢舉]  
回答時間：2006-04-19 09:19:26

He does not have a stable income, competition is fierce at work yet the sales pressure is high. At times he has to deal with some very rude customers.

Figure 5-2 A screenshot of ‘Yahoo Knowledge’

The above thread in the format of question and answer is a typical example of the results generated in online discussion boards such as Yahoo Knowledge. This kind of result does not seem to be representative of how the language point in question is typically used in the target language. Nonetheless, the study found several instances of such search. The small sample (N=17) in this dataset precluded the possibility of generalizing this phenomenon to a wider population of EFL learners in Taiwan, however, it did shed light on the language learning resources of which the learners availed themselves, in addition to the more common ones anticipated (e.g., printed, electronic or online dictionaries). It has some bearing on FL pedagogy: with the advent of computer technology and accessibility of the Internet, learners nowadays may have far more resources for language learning at their disposal than teachers or researchers give them credit for. It is therefore important for teachers to explore more



possibilities for technology-enhanced learning, in order to guide learners to make extensive and comprehensive use of various resources, but at the same time to sensitize them to the potential pitfalls and/or limitations of such resources.

This section has provided an overview of the referencing behaviour identified in the behaviouristic data (i.e., query records). The mental processes underlying the observable behaviour were reconstructed through a synchronization of behaviouristic data and mentalistic data (i.e., concurrent verbal reports). The mental processes taking place as learners approached the DALC task are detailed in Sections 5.2 and 5.3.

## 5.2 Cognitive strategies

As noted in Section 2.5.2, three previous studies, Kennedy and Miceli (2001, 2010) and Sun (2003), have looked into the processes during corpus consultation, but none of them employed both behaviouristic and mentalistic measures to reconstruct the cognitive processes taking place within learners' mind as they undertake such an endeavour. To understand how learners cognitively manipulated corpus data in the DALC task, I examined their concurrent verbal reports and parallel corpus queries to infer the thinking processes as they approached the task. Drawing on O'Malley and Chamot's (1990) taxonomy of LLSs and Sun's (ibid.) cognitive skills for concordancing (see 2.4.1.2 and 2.5.2 respectively) as the preliminary coding scheme of verbal protocols, this study identified a diverse array of *cognitive* and *metacognitive* strategies employed by the learners to fare with the DALC task. *Cognitive processing* is concerned with the operative functions which involve direct manipulation or transformation of learning materials (Brown and Palincsar, 1982), as opposed to the executive functions of *metacognitive processing*, such as planning of learning, monitoring language comprehension or production and evaluating learning process and product (O'Malley et al., 1985). As the DALC task was highly specific in nature, in terms of the target language point (i.e., collocation) and the task demands (e.g., consulting corpus resources), not all the strategies in the preliminary coding scheme fit well with the mental processes found in the current dataset. The coding scheme was modified in the light of the present findings.

For the purposes of reporting, the thinking processes are discussed atomically in the form of discrete strategies, however, it is imperative to acknowledge that they are by

no means distinct and clear-cut, as will be presented in the following discussion. The thinking processes taking place in the course of corpus consultation are not so much *linear* as *discursive*. The present study identified and described instances of strategy use, but did not attempt to generalize a model to account for all corpus consultation endeavours. As McEnery and Wilson (1997:6) put it, ‘even with the most careful set of instructions, a key feature of corpus-based learning is divergence: different students take different paths through the data and find slightly different things’. Corpus consultation is highly idiosyncratic, depending on many external factors (e.g., task demands, tools consulted, search results) as well as internal ones (e.g., prior knowledge, individual differences, how language input is processed), and thus cannot be reduced to any simplistic, universal pattern. Instead of generalizing a common pattern, this inquiry into the thinking processes underlying corpus consultation behaviour aimed at accounting for the changes that occurred in the learners’ collocational knowledge (as shown in the last chapter) as a consequence of DALC. It has been claimed that more intense processing of information input leads to more effective learning (see 2.4.1.2 for the *depth of processing hypothesis* and the *involvement load hypothesis*). The previous chapter has illustrated that DALC facilitated collocation learning, and a further probe into the underlying processes may shed light on the links between the product and process of such learning.

The following subsections provide a descriptive account of how learners fared with the DALC task with *cognitive strategies*: that is, how mental capacities interacted directly with corpus evidence. O’Malley and Chamot (1990:138) characterize cognitive strategies as involving ‘interacting with the material to be learned, manipulating the material mentally or physically, or applying a specific technique to a learning task’. The definitions of the cognitive strategies identified in the present dataset are listed as follows:



**Table 5-1 Definitions of the cognitive strategies identified in this study**

Cognitive strategy	Definition
Deduction/ Induction	Consciously applying learned or self-developed rules to produce or understand the target language.
Grouping	Ordering, classifying, or labelling material used in a language task based on common attributes.
Inferencing	Using available information to guess the meanings or usage of unfamiliar language items associated with a language task, to predict outcomes, or to fill in missing information.
Substitution	Selecting alternative approaches, revised plans, or different words or phrases to accomplish a language task.
Translation	Rendering ideas from one language to another in a relatively verbatim manner.

Adapted from O'Malley and Chamot (ibid.)

### 5.2.1 Deduction

While O'Malley and Chamot (ibid.) classify the strategies of *deduction* and *induction* under the same category (see Table 5-1 for the definition), the present study distinguishes between these two different processes of reasoning. *Deduction* is viewed as synonymous with testing linguistic hypotheses in the present study: for example, consulting concordances to verify the collocability of a word combination the learner has in mind. *Induction*, on the other hand, is the reasoning process by which the learner generalizes a target collocation from instances of language in use, because s/he has no prior knowledge of the collocation. This subsection focuses on the *deduction* strategy and leaves the *induction* strategy for the next one (5.2.2).

The following is an instance of learner S5 employing the deduction strategy to verify the candidate collocations she had in mind. S5 made the observation below as she examined Question 3-1 on the worksheet (i.e., city = u\_\_\_\_\_ area; country = r\_\_\_\_\_ area, see Appendix 10 for the worksheet):

**Table 5-2 A segment of S5's verbal protocols and parallel queries**

Tool and query	Verbal protocol
NTNU [urban area] [rural area]	City... I think that it's... it's a city. And it (the question) says what area... then I think it should be the area of a city. City equals... 'urban'... 'urban'... I think it's... the area of a city. [...] And then country (said in English) is country, country... I think it's 'rural' because I've learned three kinds before, 'urban', 'suburban' and 'rural'.

S5 queried [urban area] and [rural area] in one of the concordancers as she made the remark above. Given that there was neither extended contextual information nor candidate collocates given in this gap-filling question, and the respondent explicitly

cited three words (i.e., *urban*, *suburban* and *rural*) from her repertoire, it is clear that the queries (i.e., *urban area* and *rural area*) were retrieved from her linguistic knowledge. However, she was uncertain about the collocability of the combinations, because the adjectives might have been learned as discrete units, independently of their noun collocates. Therefore, she resorted to concordancers to test her hypothesis about the collocability of the candidate combinations. The hypothesis was confirmed by the exact instances of the queried combinations in concordances.

Another interesting issue arises from this episode: whether the words the learner had in mind (i.e., *urban*, *suburban* and *rural*) were learned and/or taught as discrete units or integral parts of collocations. S5 responded as follows:

*Usually I learn the words only, 'rural', 'urban'... Because when 'urban' occurs, the teacher will add its antonyms, so they come up together. I forgot whether they were followed by 'area', because I only memorized the preceding words. I didn't memorize... Because I already know their meanings, 'rural' or 'urban', they must be followed by some kind of 'area'.*

It seems that the words had been taught along with other paradigmatically related words (e.g., synonyms, antonyms), but not the syntagmatically related ones (i.e., collocations). As shown in the respondent's reflection, she seemed to assume that the frequent collocations of *rural* and *urban* did not need to be taught or learned together with the adjectives themselves, because she 'already know their meanings', implying that collocates could be inferred solely on semantic grounds. The learner's assumption shows a lack of awareness of collocational restrictions. Vocabulary pedagogy in FL contexts needs to raise learners' awareness of the syntagmatic relations of words in addition to the paradigmatic ones, as both constitute integral parts of word knowledge.

In addition to testing hypothesis about collocations, participants used concordancers in unexpected yet innovative ways. Question 2-1 asked the learner to choose one miscollocation from four options:

Q2-1: highly educated; highly exhausted; highly profitable; highly unusual.

S2 might have mistaken the adjective *unusual* for the adverb *usually*, so she suspected that the combination *highly unusual* was incorrect because the other three options were grammatically well-formed (i.e., adverb-adjective). Her misapprehension is evidenced in the following remark:

*If it (unusual) does not belong to this word class (adjectives), it's not supposed to be preceded by an adverb.*

She seemed to assume that *unusual* was not an adjective, so should not be preceded by an adverb. Therefore, instead of searching for the exact collocation *highly unusual* in the concordances, she queried to examine the words preceding *unusual* in order to determine its word class. Contrary to her conjecture, the concordances showed many instances of *unusual* pre-modified by adverbs.

*'Unusual'... and it displays many instances. And then... check if it (unusual) is preceded by adverbs. So... it seems... my guess seems to be wrong. Ha ha, because it seems to be an adjective, too.*

The concordances provided counterevidence in rejecting her hypothesis and convinced her that *unusual* was an adjective rather than the frequency adverb *usually* for which she had mistaken.

Overall, in the first instance, S5 used corpus data to verify her hypothesis of *collocation*, while S2 observed concordances to test the hypothesis of the *syntactic category of a word*. The two instances above have shown how learners employed the *deduction* strategy along with corpus resources to approach the DALC task.

### 5.2.2 Induction

As opposed to the *deduction* strategy noted in the previous subsection (5.2.1), *induction* is a bottom-up reasoning process in which the learner does not have any linguistic hypothesis in mind to be tested, but generalizes a language pattern from instances of language in use. The induction process is basically what the DALC task aims to stimulate: the learner generalizing collocations through an observation and analysis of real language data in corpora. In the course of induction many other strategies may also come into play, such as grouping or differentiating language data, or inferencing from various knowledge sources. In this subsection, I will illustrate an overall induction process and then move onto more specific strategies used as an integral part of induction.

In the example below, S15 was asked to induce another verb collocate of the noun *method*:

Q3-2: to use a/the method = to e\_\_\_\_\_ a/the method.



Table 5-3 shows a segment of the learner’s verbal reports and parallel corpus queries. The alignment of the two datasets allowed me to reconstruct S15’s induction process: she first used the *POS search* (see query methods in 5.1.1) querying the noun *method* for verb collocates, but did not find any with an initial *e*; so she decided to turn to another concordancer to try a *node search*, querying only the node word *method*; she tried with little success until consulting the third concordancer with another *POS search*. Among other verb collocates, she located two that might be the answer to this question, *employ* and *explain*, as they both had an initial *e*. Having difficulty choosing between the two, she consulted an online dictionary, and decided on the verb *employ*, as one of its meaning senses was ‘to use’, which seemed to match the intended meaning of the question.

Table 5-3 A segment of S15’s verbal protocols and parallel queries

Tool and query	Verbal protocol
Collins [VERB+the+method]	<i>It (the question) means looking for another... ‘use a method’... ‘use a method’, another verb for ‘use’? Let me see. [...] Verbs followed by ‘method’. Found ‘use’, it’s already there in the question. With an initial ‘e’... Doesn’t seem to have anything that starts with ‘e’. And then change (reference tools).</i>
NTNU [method]	<i>I’m looking for the verb, doesn’t seem to have any here.</i>
IWiLL [verb+method]	<i>See if there’s that verb search (function). [...] Found ‘employ’... but... I saw ‘employ’, but I should check the meaning to see if the meaning of the sample sentence matches. Not quite sure. [...] Saw ‘explain’, but it means to explain, not exactly the same.</i>
Yahoo Dictionary [employ]	<i>I’ll check its (employ) meaning to see if there’s any other meaning that I haven’t learned before. [...] It means to hire, use, make use of. I feel that ‘use’ or ‘make use of’ can go with ‘method’. ‘Use this method’, so I think it should be e-m-p-l-o... ‘employ’.</i>

Note: VERB is the part-of-speech tag for verbs in Collins WordbanksOnline English Corpus.

In this segment, the learner used three concordancers to look for language evidence for a collocation she had not known by the time of the task. She was successful in inducing the target collocation by alternating query methods and browsing through the language data provided by various corpora. Not only did she generalize the target collocate from corpora, but she also consulted an online dictionary to confirm the word meaning of the collocate. Induction processes, as demonstrated above, were very common in the current dataset, because they were what the DALC task set out to stimulate – having learners exposed to multiple instances of how a word is typically used and eventually being able to induce frequent collocations from a plethora of language evidence. It is widely recognized that corpus consultation has the potential for ‘the development of abilities to “identify-classify-generalize” on the basis of

language experience, one of the abilities on which learning in general, and autonomous learning in non-institutional settings in particular would seem to rely' (Bernardini, 2004:17). As highlighted in Chapter 2 (2.3.2.3), at the heart of DALC is the learner's role as a researcher to observe, analyze and interpret patterns of language in use, as evidenced in the example above. The induction processes, albeit laborious, may go a long way toward autonomous learning. The subsections that follow elaborate on more specific strategies that contribute to the induction processes: grouping and differentiating (5.2.2.1) and inferencing (5.2.2.2).

### 5.2.2.1 Grouping and differentiating

The processes of *grouping* and *differentiating* were often found to be alternated as a means to distinguish a target form from other candidates. The following is an example of the alternating use of *grouping* and *differentiating* strategies as S4 approached Question 1-2:

- Q1-2: India's \_\_\_\_\_ economy has brought huge numbers of foreigners and newly affluent (富有的) Indians to the country's big cities (Choose ONE unacceptable option).  
(A) booming (B) extending (C) growing

Table 5-4 is a segment of the learner's verbal protocols aligned with the reference tool she consulted, the queries entered and the results generated as she approached the question above. At the outset of this segment, S4 initially distinguished the combination *booming economy* from the other two candidates (i.e., *\*extending economy* and *growing economy*) based on her knowledge of the meanings of the words *extending* and *growing*. She then supported her argument by making reference to a contextual clue, the word *affluent*, which seemed to her cohesive with the assumed meanings of *\*extending economy* and *growing economy*. However, her preliminary conjecture was challenged after finding out in an online dictionary the meaning of the word *boom*. Her grasp of this word led her to think that the three options were semantically similar. She grouped the three candidate collocates on a semantic basis, but was then faced with the challenge to differentiate them in order to answer the question. S4's earlier attempt to approach the question from word meanings failed, so she resorted to the collocates of the option words as a means of distinguishing them. She redirected her attention to the collocates, and concluded that *boom* collocated with things such as *business* and *distance*, whereas *extend* was associated with *time* and *body movement*. Irrespective of the accuracy of her

observation, the range of collocates she observed seemed to be distinctive enough for her to differentiate the miscollocation *\*extending economy* from the other two options.

The alternating processes of *grouping* and *differentiating* (i.e., compare and contrast) were evident in this example: first, singling out *booming economy* from the other two combinations based on partial comprehension of word meanings, then grouping three candidates according to word meanings, and finally distinguishing *\*extending economy* from other combinations judging from the respective ranges of collocates. Though the learner did not make use of corpus data, one of the key objectives of the DALC task was achieved in this episode: that is, raising collocational awareness. This was evident in the change in the ways the learner judged collocability, from dwelling on word meanings to observing the range of collocates. The transition was desirable, as the learner seemed to be sensitized to collocational restrictions, in the sense that collocability was not always based on semantic grounds.

### 5.2.2.2 Inferencing

Inferencing is a common strategy employed by the learners to bridge the gap in various aspects of linguistic knowledge. Haastrup (1991:40) regards inferencing as a process that ‘involves making informed guesses as to the meaning of an utterance in light of all available linguistic cues in combination with the learner's general knowledge of the world, her awareness of context and her relevant linguistic knowledge’. The three major knowledge sources of inferencing pointed out by Haastrup (ibid.): *contextual information*, *linguistic knowledge* and *world knowledge* were all found to be exploited by the learners in the present study.

#### 1) Contextual information

It is not unusual for learners to draw on textual contexts in tests or worksheets to infer unknown information. However, few instances of inferring from textual contexts were found in the present study, owing to the question types in the DALC worksheet. The worksheet (see Appendix 10) consisted of MC questions with and without contexts, gap-filling questions, error identification and correction questions, semantic grid and translation. None of the question types had extended contexts from which learners



Table 5-4 A segment of S4's verbal protocols and parallel queries

Tool and query	Search result	Verbal protocol
Yahoo Dictionary [boom]	<p>n. (名詞 noun)[C]            (雷,大砲等的)隆隆聲;(波浪的)澎湃聲;(蜂等的)嗡嗡聲            (商業等的)景氣,繁榮;(政治形勢等的)突然好轉            The city is having a building boom.            這個城市正在大興土木。            激增,暴漲</p> <p>vi. (不及物動詞 intransitive verb)            發出隆隆聲(或嗡嗡聲等)            The big guns boomed in the distance            遠處大砲聲隆隆作響。            聲音隆隆而過            激增,暴漲;迅速發展,興旺            The country boomed when gold was discovered there.            那個縣在發現了黃金後便繁榮起來。            Business boomed after the depression.            蕭條之後生意興隆起來。            名聲大噪</p> <p>vt. (及物動詞 transitive verb)            用隆隆聲表達[(+out)]            The foghorn boomed out its warning.            霧角鳴聲示警。            使迅速發展;使興旺;使出名            Foreign investments boomed the city.            國外投資使這個城市迅速繁榮起來。            為...大力鼓吹</p>	<p><i>I think it (the answer) might be A. As far as I know, both 'extending' and 'grow' have the meaning of... this one (extending) is expanding and that one (growing) is boosting. They both make sense with 'economy', boosting economy or economy expands freely. It (the question) says 'affluent' here, so (economy) should be going upward. This one (booming)... don't know the meaning. [...] I found that it means... sudden increase, so the meanings of the three (options) are similar. Then I will look at (the words) that collocate with (booming). [...] I saw it (booming) has to do with ... distance and also with business. [...] Uh... I think it (booming) is also possible, because economy and business are similar... have something in common.</i></p>
	<p>vt. (及物動詞 transitive verb)            延長,延伸;擴大,擴展            Can't you extend your visit for a few days?            你們訪問的時間不能延長幾天嗎?            伸,伸出            He extended his hand in greeting.            他伸出手來表示歡迎。            致;給予,提供[(+to)]            I would like to extend a warm welcome to our visitors.            我想對我們的來訪者表示熱烈的歡迎。            使竭盡全力[H]</p> <p>vi. (不及物動詞 intransitive verb)            伸展;擴大;延續            The tourist season extends from May till October.            旅遊季節從五月延續到十月。</p>	
Yahoo Dictionary [extend]		<p><i>So I'm going back to check 'extending'. I saw that 'extend' seemed to go with... time or movements, body movements. Yeah, so I think the answer is not 'extend' because... economy is an abstract thing and has little to do with time and movements, so I think (the answer) should be B.</i></p>

may infer unknown words and/or collocations. The question types, to some extent, inhibited the use of textual contexts as a source of inferencing. The following is an episode of a learner inferring adjective collocates of the noun *issue* in the question below:

Q3-4: Death penalty is a c\_\_\_\_\_ issue. Some argue that it violates human rights, while others view it as a necessary evil to prevent crimes.

In the segment below, the learner S9 was trying to make sense of the question using textual information in combination with her world knowledge:

*S9: My thought is... the question is about... a very... very serious issue or something like that. Because... because it says here that 'some argue that is'... something about violence... and human rights. Anyway there is a quarrel, so it's supposed to be... uh... arguing over something, so I suppose this is a very serious topic or something.*

[S9 found 'critical' in concordances and consulted Yahoo Dictionary for its definition]

*S9: Because it (critical) means critical and judgmental, I think issues are often like that... I mean... whenever 'issue' occurs, it's got that kind of feel.*

*R: Are you saying that 'issue' goes with it... with 'critical'?*

*S9: Yeah, yeah, yeah... it's what happens when you watch too much politics (on news), haha...*

In the first half of this segment, S9 drew on contextual information, inferring that there was an argument over some important issues, from key words *argue*, *violence* and *human rights*. However, after consulting the meaning of the word *critical*, the core meaning sense thereof (i.e., being inclined to criticize severely and unfavourably) seemed to activate her world/schematic knowledge of the discourse in politics. This example was a combinatory use of strategies in which contextual information and the learner's world/schematic knowledge both contributed to the inferencing process. The textual context helped the learner to demarcate the semantic field of the target collocate, and a core meaning sense of the candidate collocate *critical* went on to activate the learner's world/schematic knowledge in which the word *issue* was strongly associated with the negative sense of *critical* and the political discourse. This instance has provided a glimpse of the complex interaction between external mediating tools (e.g., textual information, language input from corpora and dictionaries) and internal cognitive processing (i.e., making sense of the text and activating world knowledge) taking place in the course of the DALC task.

## 2) Linguistic knowledge

Given that the language point at hand (i.e., collocation) was a component of lexical knowledge, it was anticipated that the learner would draw on various aspects thereof as sources of inferencing (see 2.2.1 for components of lexical knowledge). Closely related to lexical knowledge is the grammatical knowledge that concerns the syntactic categories of words and the orders in which words occur in a sentence. These two aspects of linguistic knowledge were found to be used by the respondents as sources of inferencing in the DALC task: *lexical* and *grammatical knowledge*.

### a) Lexical Knowledge

While lexical knowledge entails several components, the respondents in this study drew primarily on three components as sources of inferring collocates: *word meanings*, *analogous collocations* and *word associations* (e.g., synonyms or antonyms).

#### i) Word meanings

It is not uncommon that a learner resorts to semantics to infer collocates of a word, since word meaning (i.e., form-meaning link) is the first and foremost aspect of lexical knowledge. In particular, for FL learners who do not have a developed repertoire of the TL in use, they may well draw on the resource that springs to mind readily, such as word meanings. Nevertheless, as noted in Chapter 2 (2.2.2.2), collocation, concerned with *word use*, is dictated by actual usage in a language, and thus does not always conform to the semantic grounds. Therefore, inferring collocates from word meanings may not always be seen as successful. In the following example, the learner resorted to the definitional knowledge of words to decide on the target collocate. The collocation question at hand is as follows:

Q4-3: Labour supply did not increase to achieve demand.

The learner was asked to identify and correct the miscollocation in the sentence above. The segment below illustrates how S1 made a collocational choice based on the definitional knowledge of words:

*I'm thinking... seems about right... there is no such thing as 'achieve demand' as far as I know, but I'll check anyway. [...] Give it a try... nope... so it means that this one (achieve demand) should be wrong. Then I would delete 'achieve' and enter V-E-R... (VERB is the POS tag for verbs in the*



concordancer Collins WordbanksOnline) *use that search method to see what goes with 'demand'. [...] I see many (verb collocates of 'demand') and I'm thinking... I think this question should be changed to 'match'. Because... it (the question) seems to mean 'meet demand', 'achieve' does have (the sense of) 'meet'. If what it (the question) means is 'meet demand', then it'd be more appropriate to use 'match', so I would rephrase this one as 'match demand'.*

Having identified the miscollocation *\*achieve demand*, S1 went on to test her hypothesis by querying this particular combination in a concordancer. The results verified her hypothesis that the verb *achieve* did not collocate with the noun *demand*. From the concordances generated she also found a range of verb collocates of *demand*. She then decided on the verb *match*, and excluded the other options, because it seemed to her that the word meaning of *match* fit well with the intended meaning of the context. It appears that she compared the words *achieve* and *match* based on her prior knowledge of their meanings, and found the word *match* a better replacement in this particular context. In this example, the learner first used the language evidence in corpora to verify her hypothesis of collocability (i.e., *deduction*) and then proceeded to look for potential collocates of the target word (i.e., *induction*). However, as shown in the segment above, it was her definitional knowledge of the candidate collocate on which she drew to make the final collocational choice. Overall, in this instance, the language evidence in corpora provided raw materials which the learner cognitively processed using the strategies of *deduction*, *induction* and *inferencing* from her prior knowledge of word meanings. Again, the cognitive processes entailed in approaching the DALC task were complexly interconnected, and by no means as clear-cut and atomized as presented here for the purposes of reporting.

Another respondent, S14, reached the same conclusion to the same question. She further justified her collocational choice by comparing the meanings of the two candidate collocates, *match* and *meet*:

*S 14: So it (demand) goes with either 'match' or 'meet', I think both are fine. But... given the literal meanings, I would choose 'match'. I would choose the more familiar one, it feels more familiar.*

*R: Isn't 'meet' (familiar)?*

*S14: And so is 'meet'. But... from what I've learned, I feel that 'meet' is often used in greetings, so I think 'match' is more suitable.*

At the outset, S14 seemed to think that both verbs, *match* and *meet*, collocated well with the noun *demand*, considering that they were both found to go with *demand* in concordances. However, having considered the word meanings of the candidates, she was inclined to choose *match*. She appeared to attend only to the core meanings of the

words, considering the meaning of *match* as 'providing something that is suitable for a particular situation or purpose', whereas taking *meet* as 'encountering someone' (e.g., in greetings such as *nice to meet you*) instead of a less core meaning sense, 'providing something to deal successfully with a situation'. It seems that the core sense of *meet* eclipsed the less marked one in the learner's knowledge, and thus rendered *meet* a less suitable collocate of *demand*.

This is a perfect example of how a learner makes collocational choice on the semantic grounds, which is a common source of inferring collocations, but also a cause for concern when it comes to those who lack an awareness of collocational restrictedness. In cases where the precise collocation is lacking, some learners may well be inclined to infer one from word meanings, as it seems to be a logical and straightforward way to express the intended meanings. As evidenced in learners' free collocation production (4.2.4), approximately half of the miscollocations found in the writing assignments (43.48% in the pre-intervention writings and 58.75% afterwards) resulted from producing word combinations on the semantic grounds (i.e., paraphrasing). Nonetheless, such attempts are not always successful given the *usage-based* and *language-specific* nature of collocations (see discussion on the arbitrariness in the choice of collocates in 2.2.2.2). For example, as illustrated in Section 4.2.4.2, a learner modified the noun *door* with the adjective *blond* (*\*blond door*) based on the word meaning of *blond*. While the adjectives *blond* and *golden* may denote the same colour, they do not share the same range of noun collocates. Many researchers warn against the over-reliance on semantics for producing word combinations (Howarth, 1998a; Pawley and Syder, 1983). This is not to claim that semantics is an invalid element as far as lexico-syntactic relation is concerned, but to recognize the unreliability of such an element when dealing with a language aspect primarily based on actual usage, such as collocation.

## ii) Analogous collocations

In addition to word meanings, learners were also found to infer collocations from the ones that they took to be semantically related. Because the nodes in such collocations were semantically related to the target node, they were assumed to share the same range of collocates. For instance, *\*pay notice* is a result of analogy from the common collocation *pay attention*, because *notice* and *attention* are similar in meaning, and

thus are mistaken to share the verb collocate *pay*. This example again illustrates that collocability is not necessarily based on the semantic grounds, but is dictated by actual usage in a language in some cases, as highlighted in the previous subsection.

The following is an example of the learner inferring a collocation from a semantically related one. In an attempt to translate the sentence below, S1 turned to analogous collocations for sources of inferencing:

Q6-1: 賺取穩定的收入並非難事 (It is not difficult to earn a stable/steady income.)

S1 consulted a corpus and an online dictionary for adjective collocates of the noun *income* and came across two candidates, *monthly* and *regular*. See the table below for the results generated and her verbal reports.

Table 5-5 A segment of S1's verbal protocols and parallel corpus queries

Tool and query	Search result	Verbal protocol
Collins [JJ+income]	...fund. If you want a regular stream of <b>monthly income</b> , you can get that from a mutual fund, plus a...	<i>I'm considering using this word (monthly) because a stable/steady income means getting salary every month, so I'll consider this word. However, I'd better check first if there are any better options. Then again, I... I saw this... 'if you want a regular stream of monthly income', I'm wondering if I could use 'regular' instead. [...] Now I see 'regular job', 'regular job', Then... if compared with 'monthly', I'd probably go for this one (regular).</i>
Yahoo Dic. [regular]	He's got no regular job. 他沒有固定的工作。	

Note: JJ is the part-of-speech tag for adjectives in Collins WordbanksOnline.

Being unable to locate the exact collocation *stable/steady income*, S1 was initially inclined to use the collocation *monthly income* found in the concordance line, because '*a stable/steady income means getting salary every month*' (S1's remark). It seems that, at this point, she attempted to paraphrase around the lack of knowledge of the exact collocation *stable/steady income*. However, having examined the same concordance line again, she came to notice the word *regular*, which she postulated to be another potential collocate. In order to decide between the two candidates, *monthly* and *regular*, she consulted the word meaning of *regular* in an online dictionary where



she found the collocation *regular job* in one of the sample sentences. The causal relationship between *job* and *income* seemed to lead her to assume that they shared the same adjective collocate *regular*. She did not actually find any instances of the combination *regular income* from corpus data. Rather, her decision on the combination *regular income* was a result of inferencing from a dictionary-attested, semantically-related collocation *regular job*. Moreover, the shared sense of constancy between *regular* and *stable/steady* may have also made the former seem like a possible substitution for the latter in this case, irrespective of the difference in Chinese translation. It is worth noting that though the final decision was made mainly on the basis of analogy, it would not have been possible if the learner's *world knowledge* had not come into play. The conceptual association between a job and the income (i.e., the money made for doing the job) that made the collocation *regular job* analogous to *regular income* in the learner's mind was the result of activating her world knowledge. This episode has demonstrated that the learner was capable of synthesizing a number of strategies and knowledge sources of inferencing (i.e., paraphrasing, inferring from analogous collocations and world knowledge) as a means of approaching the DALC task. This example lends support to Nassaji's (2003) argument that successful inferencing is not so much the result of using one particular strategy or knowledge source but converging and linking various kinds of strategies and knowledge sources.

A number of other instances of inferring from analogous collocations were also found in the current dataset, including *stable income* inferred from *stable position*, *break a/the contract* from *break one's engagement*, and *achieve power* from *achieve fame*. The first instance (i.e., *stable income* from *stable position*) which entailed a combination of two knowledge sources, analogous collocation and world knowledge, will be discussed at the end of Section 5.2.2.2.

### iii) Word associations (antonyms)

The respondents also drew on word associations to make inferences. While instances of drawing on word associations such as *synonyms* and *antonyms* were both found in the dataset, the use of the former was categorized under the strategy of *substitution* because *synonyms* were found to replace the target collocates, whereas the latter was identified as one of the knowledge sources of the strategy of *inferencing* because *antonyms* were used to infer rather than replace collocates. Hence, this subsection

discusses the instance in which antonyms served as a knowledge source of inferring collocates, while substitution of collocates with synonyms will be addressed in Section 5.2.3.2.

In the following segment, S17 was required to identify and correct the miscollocation in the sentence below:

Q4-4: The pressure grows up for her as the mid-term exam approaches.

Based on her prior knowledge of the phrasal verb *grow up*, S17 successfully identified the mismatch between *pressure* and *grow up*:

*'Grow up' (said in English) is usually used as in 'grow up' or 'growth'... rarely used to say 'pressure increases'. So I'm gonna check 'pressure'. [...] I wanna find... find something... that has to do with... a word that has to do with becoming larger in amount. [...] I saw a 'decrease', I suppose it's fine to use its opposite, 'increase', to express (the meaning of) increase.*

S17 had a grasp on the intended meaning of an increase in pressure, but she was unable to locate any collocations carrying the meaning. However, she then found a collocation *pressure decreases* in concordances, and thus inferred the collocation *pressure increases* based on her knowledge of the paradigmatic relation between the two verbs, *increase* and *decrease*.

This subsection has illustrated how respondents inferred collocations from various aspects of lexical knowledge, including *word meaning*, *analogous collocation* and *word association*. The next subsection moves to another aspect of linguistic knowledge, grammatical knowledge, to demonstrate how it contributed to inferences.

#### **b) Grammatical knowledge**

In some cases learners drew on their knowledge of the syntactic categories (e.g., verbs, adjectives or adverbs) of words and knowledge of orders in which words occurred to help them approach the collocation questions. The following example illustrates how grammatical knowledge contributed to the identification of a miscollocation.

The question below asked the learner to identify and correct the miscollocation in the sentence:

Q4-3: Labour supply did not increase to achieve demand.

In the segment below, S11 successfully identified the miscollocation *\*achieve demand*, and she justified her decision as follows:

*As for Question 3, I'd check to see if 'achieve demand' is correct. Because this 'labour supply' is the subject, subject followed by an auxiliary verb and then a verb, it must be correct. And then (another) verb... two verbs preceded by 'to' is also correct. Collocations are usually verb-noun or adjectives and something like that... or adverbs, so I think 'achieve demand' is most likely to be wrong.*

She approached the question by analyzing the syntactic structure of the sentence and reached the conclusion that it was grammatically well-formed. Given that the remaining part of the sentence was syntactically correct, it seemed to her that the verb-noun combination at the end was suspicious. Also, the co-occurrence of a verb and a noun in the combination *\*achieve demand* appeared to fit well with her knowledge of what commonly constituted a collocation. Therefore, she identified the combination *\*achieve demand* as the miscollocation and proceeded to search for replacements.

This subsection has given a descriptive account of how respondents drew on various aspects of linguistic knowledge (e.g., lexical and grammatical knowledge) to infer target collocations. In the next subsection, I present examples in which learners went beyond linguistic knowledge to world/schematic knowledge for sources of collocational inferencing.

### 3) World/schematic knowledge

World knowledge is a common source of inferencing (cf. De Bot, Paribakht and Wesche, 1997; Nassaji, 2003) when the linguistic knowledge needed is lacking. It can be used for comprehension as well as production in the course of a language task. The following episode illustrates how a learner inferred a collocation from her world/schematic knowledge of the concept, *work*.

Q 6-1: 賺取穩定的收入並非難事。(It is not difficult to earn a stable/steady income.)

To answer the above translation question, S11 translated most parts of this sentence without assistance, except for the adjective collocates of the noun *income*, which she had difficulty translating and thus resorted to corpus resources. To search for the adjective collocates of *income*, she queried the adjective in Chinese [穩定的]



(stable/steady) on a bilingual concordancer and came across the concordance line below:

Later, after women attained a more stable position in the workplace, we could make our own choices as to how to dress.

而後女人在工作場合有了較穩定的地位，  
她便可以自行做選擇

Though there was not the exact collocation *stable income* in the concordance line, S11 noticed the collocation *stable position* therein. She assumed that *stable* could collocate

with *income* because of the attested example of *stable position*. She explained her collocational decision on *stable income*:

*I checked on the question and it is about 'stable income', so... um... to do a search... It is reasonable to look for something related to status... or money. So... let me see... [...] I see... social status, peace, important orders or problems, and so on. So none of these fits... but I just saw one (instance)... um... this sample sentence says 'more stable position in the workplace', so I feel that has something to do with money, so I think it (the collocate) should be 'stable'.*

It seems that S11 associated *income*, namely the money earned from work, with the *position in the workplace* based on her world/schematic knowledge. She may have assumed that *income* and *position in the workplace* were conceptually related in the sense that they both fell in the domain of the concept 'work'. The two words therefore shared the same collocate *stable*. The rationale behind her choice of collocate was evidently a conceptual/schematic one motivated by her world knowledge of the conceptual association between *income*, *position* and *work*.

The next example is a failed attempt at inferring a collocation from world knowledge. Question 4-2 asked the learner to identify and correct the miscollocation in the sentence:

Q 4-2: She was about to step on a major expedition.

Having examined the question, S11 successfully identified the miscollocation \**step on a major expedition*, and went on to search for acceptable verb collocates of the noun *expedition*. She had come across a number of candidates such as *join*, *lead*, *make* and *undertake*, but was later misled by the following concordance line:

...Nepal, and spent last winter on a climbing expedition in Argentina and Chile. A vivacious Scot...

Misled by the concordance line, S11 replaced the phrasal verb *step on* with the verb *climb*, and came up with yet another miscollocation *\*climb a major expedition*. Note that *climbing expedition* in the concordance was a compound noun in which the word *climbing* did not function as a gerund, but denoted the nature of this particular activity. However, S11 mistook *climbing* as a verb collocate of *expedition*. She seemed to associate the activity *mountain-climbing* with the concept of *expedition* based on her world/schematic knowledge: the former was one kind of the latter. Semantically, the word *mountain-climbing* is a hyponym of *expedition*, among other activities, and *climbing* is arguably related to the activity of *expedition*, however, the combination *\*climb an expedition* is not only semantically problematic but also collocationally unacceptable. S11 justified her collocational decision:

*I think the most likely one is 'climbing'. Since it (the question) uses 'step on', meaning that she wants to... um... she wants to stand... stand on the expedition's major... on the major expedition. So if I use 'climbing', that means... that means wanting to... accomplish the expedition. So I think it (the collocate) should be 'climbing', the verb 'climbing' can be used.*

Another possible explanation of the learner's erroneous collocational choice may well be that *climb* and the distracter collocate *step on* were similar in that they were both concrete physical movements, as opposed to the other candidates (i.e., *join*, *lead*, *make* and *undertake*). The learner might have been misled by this shared sense between *climb* and *step on*.

In sum, this section (5.2.2) has demonstrated that in the course of inducing collocations from corpus resources, learners may *group* or *differentiate* language data, or infer unknown words and/or collocations from a range of knowledge sources, including *contextual information*, *linguistic knowledge* (e.g., lexical and grammatical knowledge) and *world/schematic knowledge*. Notably, the diverse array of knowledge sources of inferencing used by the learners was indicative of their resourcefulness in deploying and synthesizing available information to make informed guesses. The finding that learners resort to linguistic as well as extralinguistic sources to tackle language problems is consistent with a body of empirical evidence (e.g., De Bot, Paribakht and Wesche, 1997; Færch, Haastrup and Phillipson, 1984; Qian, 2005).

### 5.2.3 Substitution

In addition to *deduction* and *induction*, the respondents also used the strategy of *substitution* in the DALC task. I will provide examples to illustrate the substitution of collocates with de-lexicalized words (5.2.3.1) and with synonyms (5.2.3.2).

#### 5.2.3.1 De-lexicalized words

One of the communication strategies identified in Dörnyei (1995:58) is the use of ‘all-purpose words’, which is another term for de-lexicalized words. Both terms denote lexical items that do not carry marked meanings outside the context of use, for example, *do*, *make*, *take* or *stuff*. Learners may substitute target collocates with de-lexicalized words because of a lack of awareness of collocational restrictions (see 2.2.2.2 for the arbitrariness in the choice of collocates) or mistaking de-lexicalized words as convenient or all-purpose substitution. The instance below is an example of substituting the target collocate with a de-lexicalized word. Question 4-1 asked the learner to identify and correct the miscollocation in the sentence:

Q 4-1: The meeting aims to attend the issue of violence and crime in America.

S4 successfully identified the miscollocation *\*attend the issue* and went on to replace the verb collocate *attend* with a de-lexicalized word *take*. Her rationale is as follows:

*Just like ‘take medicine’, sometimes we assume it (the collocate of ‘medicine’) is another... another verb, but it turns out to be ‘take’. Because ‘take’ goes with a lot of stuff, and it kind of... makes sense. [...] Yeah, it’s kind of multipurpose, so I suppose it (the collocate of ‘issue’) can be ‘take’. Also, it (take) has many meanings, so I think ‘take’ is fine here.*

It is clear from the segment that S4 regarded the verb *take* as a ‘multipurpose’ word which did not have marked meanings attached to it and thus could collocate rather freely. To make her case, she justified her decision with the collocation *take medicine*. In Chinese, the verb collocate of the noun *medicine* is *eat*, so the literal translation would be *\*eat medicine*, which is a miscollocation in English. The learner was aware that in English the verb collocate of *medicine* was *take*, but she did not seem to know that the verb *take* in this case was used in the sense of ‘consumption’. She mistakenly assumed that *take* was used instead of a more specific verb *eat* was because *take* was a de-lexicalized word which collocated rather freely with many nouns. Despite her false analogy, it is clear from her statement that the decision to substitute the verb



collocate with *take* was based on her assumption that it was de-lexicalized. Although de-lexicalized words in themselves carry little meaning and even gain their meanings from collocates in some cases, substituting collocates with de-lexicalized words may not always be successful because collocations are largely usage-based. Generally speaking, despite the false analogy, implicit in this segment is the learner's *metalinguistic awareness* that some words are 'multipurpose' because they collocate freely with a wide range of words.

### 5.2.3.2 Synonyms

In some cases, learners substituted the target collocate with synonyms. They would consult online bilingual dictionaries for the synonyms of the target collocate, for example:

Q3-2: to use a/the method = to e\_\_\_\_\_ a/the method

Instead of looking for the verb collocates of the noun *method* as expected, a few participants approached the question by searching for the synonyms of the verb *use*. Some participants looked up *use* in online bilingual dictionaries, hoping to find its synonyms with an initial *e*, while others went further to combine the strategy of translation, translating *use* into Chinese 使用 and then consulted online dictionaries for English equivalents. One of the synonyms of *use* found in online dictionaries was *exercise (a right)*, so those who used this approach all mistook *exercise* as the target collocate and came up with a miscollocation *\*exercise a/the method*. The collocation *exercise a right* is an example of how a word may gain part of its meaning from its collocate, because the verb *exercise* only assumes the sense of *use* when it collocates with nouns such as *rights*, *power* or *influence* in a formal genre/register. However, the participants seemed to be oblivious to its collocates and the genre/register in which it was appropriate.

Of the six participants who answered this question, three came up with the miscollocation *\*exercise a/the method* as they all resorted to the same approach described above. The other three participants found the preferred response *employ a/the method*, because they searched directly on concordancers for the verb collocates of *method* instead of the synonyms of *use*. This is not to claim that certain search

methods and/or reference tools are preferable to others, because the usefulness thereof depends not so much on their inherent nature, but on how well they lend themselves to the task at hand. The discussion here recognizes how resourceful learners can be in the exploitation of diverse language learning resources, while at the same time warns against the potential pitfalls of these tools when used injudiciously.

#### 5.2.4 Translation

Translation is a very common strategy employed by the respondents. The results show a number of instances in which translation was used for comprehension or production. The DALC task entailed the use of reference tools to approach the questions, so translation was often found to be used in combination with other strategies in the course of the task. It is nonetheless worth noting that the *usage-based* and *language-specific* nature of collocation renders it difficult to rely exclusively on translation to produce collocations (see 2.2.2.2), so such attempts were not always successful.

There were several instances in which participants looked for the English equivalents of a Chinese word to collocate with the target node word, instead of searching directly for the collocates in English. For example, the learner was asked to identify and correct the miscollocation in the sentence below:

Q 4-1: The meeting aims to attend the issue of violence and crime in America.

Having identified the miscollocation *\*attend the issue*, S13 went on to look for verb collocates of the noun *issue*. Instead of searching for the verb collocates of *issue*, she queried [參與] (*attend*, in the sense of being present at an event or activity) to look for its English equivalents. She seemed to be misled by the core sense of *attend* (i.e., being present at an event or activity) and attempted to substitute it by synonyms. She did not have any synonyms of *attend* at her disposal, so she translated it into Chinese and used the referencing tools for English equivalents. As can be predicted, none of the English equivalents she found in online bilingual dictionaries (i.e. *join*, *participate in* or *take part in*) collocated with *issue*, because the combinations would have been semantically problematic as well as collocationally unacceptable. Another instance in which translation was used in conjunction with the strategy of substituting collocates

with synonyms has been discussed in Section 5.2.3.2.

In addition to translating words or collocations, some participants used online translation tools to approach translation questions. In considering the possible ways to translate the Chinese sentence below, S3 was inclined to use online translation tools:

Q 6-2: 達不到課程要求的學生將無法畢業。

(Students who do not meet/fulfil the requirements of the course will not be allowed to graduate.)

*Ok, there are several ways, for example, um... the student who (said in English)... students... what kind of students... what kind... what kind of students cannot graduate... and then use an adjective to modify it... This is one way of translating. Alternatively, these students cannot graduate because of what... what kind of reason... So we can approach (the question) with these two ways. Actually, the easiest way would be entering Chinese (into the translation tool) for translation.*

In the segment above, S3 seemed to have in mind two syntactic structures into which the lexical items or chunks provided by the machine translation could be filled:

1. Students + relative clause + cannot graduate.
2. These students cannot graduate because + clause.

This exemplifies the operation of Sinclair's (1991) *open-choice principle* (or slot-and-filler principle) in language production (see 2.1). The syntactic structures S3 had in mind contained slots to be filled in with lexical items or chunks denoting the intended meaning of the question, in order to form a grammatical and meaningful sentence. It is clear from the blanks left that he did not have the target collocation *meet/fulfil the requirements of the course* in his repertoire. Therefore, S3 used an online translation tool to translate the whole sentence first, and then rearranged the machine-translated chunks to form a grammatical sentence:

Yahoo Translate: Cannot achieve the student who the curriculum requests to be unable to graduate.

S3 answer: The students who cannot achieve the curriculum request are unable to graduate.

The machine-translated version is clearly a word-for-word, ungrammatical one. Aware of the limitations of this reference tool, S3 used it only insofar as to obtain the lexical items and/or chunks which could later serve as the building blocks to construct the sentence. However, the *open-choice principle* seems to be successful at the sentence level, but much less so when it comes to collocations. As shown above, although the learner's answer was grammatically well-formed, the combination *\*achieve the curriculum request* borrowed from machine translation was still an unacceptable one. The credulous participant accepted the chunk provided by the



reference tool without careful examination of the collocability. The two examples above have shown that literal, word-for-word translation from another language may not be reliable in the case of *language-specific* linguistic phenomena such as collocation. Also, they have implications for language pedagogy that learners need to be guided to use reference tools and verify the results judiciously.

This section (5.2) has illustrated how the respondents deployed a diverse array of *cognitive strategies* to manipulate language data generated from various reference resources. Notably, three main cognitive strategies (i.e., deduction, induction and translation) coincide with Kennedy and Miceli's (2010) major functions of corpus consultation (see 2.5.2): *pattern-defining* is *deduction* or hypothesis-testing in the present study; *induction* is similar with *pattern-hunting* in the sense of browsing through concordances for recurring patterns; and finding equivalents in the TL for a particular L1 pattern is not uncommon in this study when learners came up with collocations in L1 and needed to translate them into the TL with the help of corpora. However, the major functions of corpus consultation in Kennedy and Miceli (ibid.) were prescribed by the researchers, whereas the strategies identified here were actually used by the learners in the course of the DALC task. Generally speaking, some strategic uses were successful, while others less so. It follows that there is no good or bad strategy use in absolute terms: strategy use varies with individual differences as well as task demands (O'Malley et al., 1985; Oxford et al., 2004), so strategies are *task dependent* or *environment dependent* (Cohen and Macaro, 2007). Bearing in mind the relative nature of strategy use, this study did not evaluate the strategies *per se*, but presented episodes to show how strategies were used, with the aim of revealing the complex mechanisms underlying the seemingly straightforward corpus consultation practice. The next section discusses the metacognitive strategies employed to coordinate the cognitive strategies described above.

### 5.3 Metacognitive strategies

In addition to the cognitive strategies employed to directly manipulate the language input obtained from reference resources, metacognition was also found in action to orchestrate cognitive processing in the course of the DALC task. As noted earlier (5.2), *cognitive processing* is the operative functions that involve direct manipulation

or transformation of learning materials, whereas *metacognitive processing* entails the executive functions that coordinate cognitive processing (Brown and Palincsar, 1982; O'Malley et al., 1985). Metacognitive experiences are, according to Flavell (1979:906), 'any conscious cognitive or affective experiences that accompany and pertain to any intellectual enterprise'. The metacognitive strategies found in the current dataset fall into three broad categories: *planning*, *monitoring* and *evaluation*.

### 5.3.1 Planning

The metacognitive strategy of *planning* is defined as 'previewing the organizing concept or principle of an anticipated learning task; proposing strategies for handling an upcoming task; generating a plan for the parts, sequence, main ideas, or language functions to be used in handling a task' (O'Malley and Chamot, 1990:137). As the DALC task usually entailed a synthesis of external mediating tools and internal mental capacities, *planning* prior to the task was commonly found in the dataset.

In the following example, S8 took some time analyzing Question 3-1 and considering the feasibility of different reference resources before he actually conducted a search.

Q 3-1: city = u\_\_\_\_\_ area; country = r\_\_\_\_\_ area.

Initially S8 had mistaken the question as one that asked for synonyms of the nouns *city* and *country*, instead of one that elicited collocates of the noun *area* as anticipated, but he later changed his mind to look for definitions of the words *city* and *country* toward the end of the planning phase:

*I'm thinking how to look for... because you want synonyms. Synonyms are difficult... uh... electronic dictionaries are easy... synonyms... can this (concordancer) generate synonyms? I'm curious. [...] Looking for the preceding or following words... synonyms... mobile phones? electronic dictionaries should be easier. This... this... let me think... I can't search backwards, that would be an even bigger hassle. [...] You can't possibly search for some kind of 'area', that will get you many stuff coming before 'area'... and you'll end up with too many (collocates) meaning 'city', isn't that even harder? [...] I'm wondering if there's anything that I can look... backwards... that equals to what kind of 'area'... But it's gonna be a huge hassle using Google. Google can't do it, I'm sure. Given I've been a long-time user of Google, I'm sure Google can't do it. It's not something Google can work out. [...] Because synonyms are usually like... um... for synonyms of 'city' I'd look for... single words that means blah blah blah, like 'city'. It (online dictionary) can't search for... um... you can't search backwards. It already hinted here that the adjective is some kind of 'area', right? But... you can't say the synonyms to be searched for... I mean... just as... um... how should I explain this? Say... take 'teacher' for example, it means 'the person who teaches students', it makes sense. However, I can't search in Chinese: what does a 'teacher' equal to. It equals to 'instructor', equals to blah blah blah. It (online bilingual dictionary) only gives you the word meanings, it hardly tells you that a teacher is 'a person who teaches students'.*

In the segment above, S8 first explored the possibilities of various reference resources for synonyms, including electronic dictionaries, concordancers, search engines and even mobile phones. It suddenly occurred to him that there might be another way of searching, 'searching backwards' in his words (possibly referring to the *node search*, see 5.1.1 for query methods), but the idea was rejected soon after it flashed across his mind, because filtering out candidate collocates in concordances seemed laborious to him. Then he began to realize that the question did not ask for synonyms, because he seemed to think that synonyms only took the form of single words. Given that the information the question aimed to elicit contained more than one word (i.e., u\_\_\_\_\_ area and r\_\_\_\_\_ area), the learner changed his mind and decided that the question asked for the definitions of *city* and *country*, instead of their synonyms. To elaborate his point, he gave an example of the word *teacher*, explaining that online bilingual dictionaries only provided Chinese translation but not definitions in English (e.g., *a person who teaches students*) which he planned to look for at this point. The planning process described above finally led him to consult an English monolingual dictionary which, according to the learner, provided elaborated definitional information of words, and therefore better served his purpose. However, he did not find what he expected from the dictionary entries. In the course of planning, S8 examined the possibilities of various resources and considered the ways in which the search could be conducted, including 'searching backwards' (i.e., *node search*), looking for synonyms of *city* and *country*, and finally the definitions thereof.

This episode has illustrated how the learner examined the task at hand, explored options and resources to handle the task, and eventually came up with the plan to approach the task. Even though the learner may have misapprehended what the question set out to elicit, namely *collocates* rather than *synonyms* or *definitions*, he was fairly strategic in formulating a plan to look for the answer to the question. As the learner's comprehension of the task demand evolved, he accommodated the proposed approaches to the task accordingly.

Wenden (1998) proposes three considerations to be taken into cognizance in task analysis at the outset of learning:

1. considering task purpose in relation to their own learning needs and goals;
2. identifying the problem type in the task as distinct from other types;
3. devising the ways to best approach the task, including the knowledge and skills needed to do so.



The considerations suggested above may provide an explanation for S8's seemingly thorough yet unsuccessful attempt: a key element was missing in the course of his planning, *considering task purpose*. Whereas he demonstrated the abilities of *problem identification* (e.g., synonyms or word meanings) and *resource selection* (e.g., concordancers, search engines or dictionaries), he misapprehended the task purpose which aimed to elicit the collocates of the noun *area*. Given that the learner misunderstood the purpose of the DALC task, it was no surprise that he did not find the answer to the question, notwithstanding a detailed planning. Since knowing the task purpose is the first and foremost element in undertaking a task, it is imperative that learners have a clear understanding thereof, in order to guide them through the language task.

### 5.3.2 Monitoring

*Monitoring* entails keeping track of the learning process and tackling problems that hinder the completion of the process with appropriate measures (Flavell, 1981). In the course of a language task, learners may consciously or unconsciously check the ongoing process and accommodate their approaches accordingly. In doing so, learners monitor the problem-solving process in order to ensure the completion of the task. Chamot, Kupper and Impink-Hernandez (1988:II-21) define *self-monitoring* as 'checking, verifying, or correcting one's comprehension or performance in the course of a language task'. In this study, *monitoring* was found at two levels: *monitoring* one's language comprehension or production, and *monitoring* the progress of the language task.

#### 5.3.2.1 Monitoring at the linguistic level

The following example is one in which the learner demonstrated the use of the monitoring strategy in the course of solving Question 4-2:

Q 4-2: She was about to step on a major expedition.

This question asked the learner to identify and correct the miscollocation in the sentence. Having identified the miscollocation *\*step on a major expedition*, S6 went on to consider which component of the miscollocation should be corrected:

*I saw... I saw... many... many words in an article. Now I'm gonna search for 'step on' in it. Ah, found it, 'step on the scale'. And the translation is 'standing on the scale' (Chinese), so my conclusion is that... 'step on' means 'stand on'. But again... I... can't find 'major expedition', so these two phrases ('step on' and 'major expedition')... these two are... the biggest bottleneck... bottleneck. [...] My difficulty is 'step on'... the meaning of 'step on' doesn't... doesn't match the question. But I couldn't find 'major expedition'... expedition... in the concordancers. So... I feel... a little bit... not knowing how to go about this.*

At the outset of this segment, S6 was uncertain about the exact meaning of the phrasal verb *step on*, because she had two meanings in mind before this segment: 1) 'taking over work' based on her prior knowledge; and 2) 'telling someone to drive a vehicle faster', drawn from the expression *step on it* found in an online dictionary. At this point she was further bewildered by a third meaning 'standing on; putting foot on something', found in an instance from an online discussion board. Neither of the three meaning senses of *step on* she had in mind seemed to match the intended meaning of the question, nor could she find instances of *major expedition* in concordances. She thus had difficulty in deciding which part of the miscollocation (i.e., *step on* or *major expedition*) was to be corrected. It was at this point that she stepped back from the problem-solving process to check her own comprehension, and identified the problems hindering her progress. She explicitly pointed out her 'bottlenecks' as 1) the mismatch between the assumed meaning of *step on* and that of the sentence; and 2) her uncertainty about the collocability of the word combination *major expedition*. Flavell (1979) argues that as learners monitor, they are prompted to examine the relationships among learning goals, task outcomes and the means to achieve this end, and to accommodate the approaches and knowledge to fit the status quo. Towards the end of this segment, the learner went beyond cognizing to metacognitively monitor her own problem-solving process by examining the relationship between the search results, the intended meaning of the question and her own comprehension, and finally identified the difficulties hindering the completion of the task. This example has demonstrated the use of *monitoring* strategy at a micro-level where the learner monitored her comprehension of a specific language point, the instance that follows will illustrate *monitoring* at a macro-level where *monitoring* was evident in the overall task management.

### 5.3.2.2 Monitoring at the level of task management

As noted earlier, in the course of a language task learners may consciously or unconsciously monitor the problem-solving process and accommodate the approaches to the task accordingly, in order to ensure the completion of the task. Therefore, fine-tuning approaches to the DALC task in the light of instantaneous corpus findings was anticipated. Also, approaches to the task may also vary with individual differences and task demands. It is the *monitoring* strategy that underlies and orchestrates the fine-tuning of approaches. While space does not permit a complete description of all the approaches taken by 17 participants, the following example gives a sequential summary of how one of the participants coordinated an array of reference resources and query methods in an attempt to answer Question 1-1:

Q1-1: It may not be easy to \_\_\_\_\_ the concept of globalization (Choose ONE unacceptable option).  
(A) grasp                      (B) seize                      (C) understand

The following descriptive account of how S1 approached the question is summarized in Table 5-6. Having glanced through the question, S1 noticed that the three options were all verbs, so she consulted a web-based concordancer that allowed for the *POS search* (see 5.1.1 for query methods), hoping to locate any of the three options. The concordances found at this point verified the collocability of *understand the concept*. She moved on to the second concordancer using the same *POS search*, but did not find any of the remaining two candidates. Then she resorted to the third concordancer which did not have a POS search function, so she used the *node search* querying the node [the concept], and found the collocation *understand the concept* again. At this point she had already consulted three concordancers with limited success, so she decided to turn to online dictionaries to look for inspirations. However, the sample sentences in dictionary entries also failed to provide the collocational information she needed. Despite the search in the online dictionary being to no avail, it suddenly occurred to her that other online resources provided by the same search engine might be useful. She then went on to experiment with an online discussion board (i.e., Yahoo Knowledge) but did not succeed, either. It was after a few failed attempts with other online resources when she decided to return to concordancers. She conducted a *node search* querying [the concept] on two other concordancers and experimented with several advanced functions with no success. It seemed to her that corpus queries



without specifying word classes were rather ineffective, so she returned to the ones that had the POS search function with a new query method, namely varying the forms of the node (e.g., *the concept*, *concept*, *concepts*). This new method did not yield expected results, but inspired her to come up with yet another new search idea, *collocate search*. Having had little success with *node search* in several reference tools, she adjusted the search methods to query the candidate collocate [grasp], and found the target collocation *grasp the concept*. Having found the evidence for the collocations *grasp/understand the concept* finally led the learner to conclude that *\*seize the concept* was an unacceptable combination. The learner's queries and the resources consulted are sequentially summarized in the table below.

**Table 5-6** *SI's queries when approaching Question 1-1*

Tool (sub-corpus)	Query
IWILL	the concept (node: verb; collocate: verb) the concept (node: noun; collocate: verb)
Collins	verb+the concept verb+the+concept VERB+the+concept
NTNU	the concept
Yahoo Dictionary	concept
Yahoo Knowledge	concept 的用法 (usage of concept) concept 的搭配词 (collocation of concept)
Yahoo Web Search	concept 的搭配词 (collocation of concept)
VLC (Language and teaching)	the concept (starts with) the concept (starts with; gapped)
Lexical Tutor (Brown)	the concept the concept (contains)
Collins	VERB+concept VERB+concepts
NTNU	grasp

A total of eight reference tools were used, including web-based concordancers, an online dictionary, a discussion board and a search engine. Query methods were constantly fine-tuned in the light of search results. The brief sequential account of the tools and methods employed has provided a glimpse into the extremely complex interplay between cognitive functioning and language input over the course of the task. In her investigation on the processes and outcomes of learners' exploration of British National Corpus (BNC), Bernardini (2000) concluded that learners tended not to look for alternative approaches when faced with obstacles. On the contrary, the present study found that when given the freedom and access to reference resources, learners could be fairly innovative and resourceful in exploiting and synthesizing resources to serve their purposes. The study provided access to a number of concordancing tools

and any resources available online, so as to resemble learners' real-life referencing endeavours as closely as possible. Even though the learners' attempts to generalize collocations from language evidence did not always turn out to be successful, they made great efforts in adjusting and experimenting with a rich array of resources and query methods.

The trial and error process demonstrated above was operated by cognitive functioning but regulated metacognitively by the *monitoring* strategy. As a consequence of *monitoring*, S1 kept accommodating to what was observed in real-time and accordingly refining her approaches to the task. She seemed to be prompted to examine the relationships among her learning goal, means of achieving it and the outcome throughout the task. *Monitoring* in the course of the task might not have been a conscious effort, but was evident in the behaviouristic data presented above: the constant refinement of the approaches to the task would not have taken place had the problem-solving process not be monitored against task specifications. Again, how learners fare with the DALC task is a highly intricate process that is neither linear nor clear-cut, several mental activities, be it cognitive or metacognitive ones, may function simultaneously with or without learners' awareness.

### 5.3.3 Evaluation

Learners may evaluate their language learning through '[c]hecking the outcomes of one's own language performance against an internal measure of completeness and accuracy; checking one's language repertoire, strategy use, or ability to perform the task at hand' (O'Malley and Chamot, 1990:137). The instance below demonstrates a learner's evaluation of her own strategy use in the course of a DALC task. The question is as follows:

Q1-1: It may not be easy to \_\_\_\_\_ the concept of globalization (Choose ONE unacceptable option).  
(A) grasp (B) seize (C) understand

S6 had consulted several concordancers before she made the following remark:

*The concordancers did not find the answer I want, so I should probably... query it (the candidate collocation) to see if it exists. [...] Ah, found it, 'grasp the concept', so it means the answer is probably right. Ok, so it means this approach works. Then, keep using this one, the same approach.*

Prior to the segment above, S6 had tried *node search* querying [the concept] or

[concept] for verb collocates with little success. So she decided to turn to *collocation search* querying the exact candidate collocation [grasp the concept] to test its collocability. Given that the collocation *grasp the concept* was found in the concordances, S6 was convinced of the effectiveness of this query method (i.e., *collocation search*), and decided to continue using it thereafter. The learner accommodated her approaches to the task in the light of search results, which was indicative of *monitoring* in action. At the same time, the successful experience made her *evaluate* the approaches taken, and finally came to the conclusion that *collocation search* was more effective. The previous section (5.2) has shown that several cognitive strategies may overlap, in this segment two metacognitive strategies also functioned simultaneously. In sum, this section (5.3) has demonstrated how learners went beyond direct manipulation of language input to *plan*, *monitor* and *evaluate* their own cognitive functioning in the course of the DALC task. The descriptive accounts provided insights into learners' intricate mental processes that coordinated internal cognition and external mediation during corpus consultation. Overall, the complex, dynamic interplay between cognitive and metacognitive functioning taking place over the course of the DALC task lends support to Schnotz's (1998 cited in Bannert and Mengelkamp, 2008:40) claim that non-linear texts impose higher metacognitive demands, compared with traditional linear texts because:

in a non-linear learning environment, a successful learner has to continuously decide where to go next and to constantly evaluate how the information retrieved is related to his or her actual learning goal. In a conventional linear text, however, the author guides the reader through the learning topic. Moreover, a hypermedia environment requires further strategic decisions by learners as the student has to choose permanently not only between various text nodes, but also between distinct information presentation formats.

The findings here have important implications for corpus pedagogy to scaffold such a cognitively and metacognitively demanding learning activity (see 7.3 for pedagogical implications). As Chamot (2004) points out, learners' cognitive styles may vary but they can all benefit from learning to use metacognitive strategies to regulate their learning.

## 5.4 Summary

Notwithstanding the widespread interests in the effectiveness of corpus consultation on language learning, a notable dearth of studies has investigated the thinking



processes underlying such seemingly straightforward endeavours. The present study employed a mentalistic measure (i.e., concurrent think-aloud) in conjunction with a behaviouristic one (i.e., corpus queries) to uncover what lay beneath the DALC task. The learners were found to be fairly resourceful and innovative in coordinating 1) physical mediating tools (e.g., web-based concordancers and online resources); 2) symbolic mediating tools (e.g., real language data in corpora); and 3) their own mental capacities (i.e., cognitive and metacognitive strategies). In addition to the web-based concordancers introduced in the intervention sessions, learners also resorted to other online resources such as dictionaries, online translation tools and even discussion boards for inspiration for collocations. As shown in the examples, learners did not simply mirror what was observed in corpora, they consciously or unconsciously activated innate capacities to process the language input generated by concordancers. Cognitive strategies employed to manipulate language input were *deduction*, *induction*, *substitution* and *translation*. In the course of *induction* learners may group or differentiate language data, or infer collocations from a wide range of knowledge sources. Beyond the level of cognitive functioning lay metacognition, which regulated the undertaking of the DALC task through *planning*, *monitoring* and *evaluation*.

As evidenced by the examples in this chapter, most corpus consultation attempts were by no means as straightforward as prescribed by previous studies (e.g., Kennedy and Miceli (2001), Sun (2003)), they were highly discursive, entailing constant changes in strategies (cognitive or metacognitive) and reference resources in the light of instantaneous findings. Moreover, various external factors (e.g., task demands, reference tools, search results) and internal ones (e.g., individual differences, prior knowledge) come into play as one consults corpus resources. Hence, corpus consultation endeavours are idiosyncratic, and cannot be easily reduced to any linear, generic model. This study thus did not generalize such cognitively complex endeavours to a simplistic model, but allowed the data to speak for itself by providing as thick a descriptive account as possible. As noted in Chapter 2 (2.4.1.2), both the *depth of processing hypothesis* and the *involvement load hypothesis* claim that more intense processing of information input leads to more effective learning. Hence, the changes taking place in the learners' collocational knowledge demonstrated in Chapter 4 can be accounted for by the immensely complex mechanisms underlying

corpus consultation illustrated in this chapter.

## Chapter 6 Examining Learner Perceptions of DALC

The previous two chapters have dealt with the *product* and *processes* of collocation learning assisted by corpus consultation, and this chapter goes on to address the third research question:

RQ3. How do EFL learners perceive the data-driven approach to learning collocations?

This chapter reports the findings on *learner perceptions* of DALC, as measured by the participants' responses to an evaluation questionnaire. Section 6.1 provides the demographic information of the respondents, followed by the English learning resources available to them in Section 6.2. As collocation is an integral part of vocabulary knowledge, the respondents' vocabulary learning experience is considered in Section 6.3. The learners' awareness of collocations is taken up in Section 6.4. Finally, Section 6.5 addresses the learners' hands-on experience and perceptions of DALC. The chapter concludes with a summary (6.6) of the results discussed above.

The questionnaire (see Appendix 13) was administered to the experimental participants at the closure of the five weekly DALC intervention sessions on 2<sup>nd</sup> April, 2009. Of the 109 participants in the experimental group, 106 participants filled out the questionnaire: 8 of the 106 questionnaires collected were regarded as invalid for missing responses to one or more of the five sections in the questionnaire. Therefore, 98 questionnaires were submitted to SPSS for descriptive and inferential statistical computations (see 3.7.3 for data analysis methods).

### 6.1 Demographic information

The first section in the questionnaire (see Appendix 13) elicited the demographic information about the respondents. Of the 98 respondents who filled out the questionnaire, approximately 87% was female and 13% was male. The profile of nationality and native language was homogenous: all respondents were Taiwanese and native speakers of Mandarin Chinese. The average length of studying English was 8.12 years, with a minimum of 4 years and a maximum of 15 years.



## 6.2 English learning resources

All participants (100%) reported having Internet access. All respondents (100%) used computers at home, approximately one-third (32.7%) of the respondents also used computers at the school computer centre, while very few did so at the school library (5.1%) or elsewhere (1.0%).

The majority (85.7%) of the informants reported having experiences of computer-assisted language learning (CALL) before, while 14.3% did not. The percentage of respondents having CALL experiences was high, probably because the term CALL was not specifically defined in the item, and thus it was open to interpretation on the part of the respondent. It could have been interpreted as strictly as exploiting specific language learning software or computer-mediated communication (e.g., instant message, email and discussion board), or as loosely as using WORD files to compose essays. A closer look at the frequency of respondents exploiting online resources to assist English learning reveals that only 5.2% did so very frequently, 26.8% answered 'often', the majority (56.7%) responded 'sometimes', while 'seldom' accounted for 10.3% of the respondents and 1.0% of the informants reported never having consulted such resources.

As for the reference tools used to aid *vocabulary learning*, only half of the respondents used conventional printed dictionaries (52.0%). The vast majority used online dictionaries and electronic dictionaries (91.8% and 80.6% respectively). The reliance on online reference tools was also evident in the think-aloud interviews: respondents favoured *Yahoo Dictionary* or *Google Translate* in particular (as discussed in 5.1.2). It was surprising to find only half of the respondents consulted printed dictionaries, given the fact that they were English majors. Many online and electronic dictionaries only provide definitions in translation, and rarely give sample sentences to demonstrate how a word is typically used, let alone frequent collocations. Such resources may suffice to provide definitional information of a word, but add little as to information of *word use* (e.g., collocations). The fact that a substantial proportion of respondents used online dictionaries calls for the need for teachers to familiarize themselves with a wider range of online resources for language learning, so as to advise learners on the strengths and limitations of these resources.

Moving to the reference tools for collocational information, 39.8% of the informants had experiences of using web-based concordancers prior to the study (see Table 6-1). The proportion was not particularly high, compared with those who had no experiences as such, but it was much higher than predicted, considering the small body of literature on learners' direct corpus consultation. It was later found that some respondents had been acquainted with one particular web-based concordancer in their writing class as a reference tool to assist writing, otherwise the proportion would have been much lower. With respect to another reference resource for collocations, namely collocation dictionaries, a considerably high proportion of respondents (85.7%) had not consulted one before (see the table below). Given the proportions of the respondents who had not consulted concordancers or collocation dictionaries, the two major sources of collocational information, there is evidence to suggest that the respondents' access to collocation reference tools was somewhat limited.

**Table 6-1** *Learner experiences of consulting reference tools for collocational information*

	Yes (%)	No (%)
Consulting web-based concordancers	39.8	60.2
Consulting collocation dictionaries	14.3	85.7

The results in this section have shown that this particular group of EFL learners had access to, and experiences of exploiting online resources for English learning, suggesting a readiness for using web-based concordancers for collocation learning. In practice, however, the learners' knowledge of available resources for consulting collocations was rather limited, a situation compounded by their reliance on online and electronic dictionaries that had little to offer as to collocational information. This provides further impetus for a wider application of DALC, in order to guide learners to avail themselves of resources for collocational information.

### **6.3 Vocabulary learning experience**

The following three subsections address the respondents' experience of learning vocabulary, including their awareness of lexico-syntactic relations of words, the perceived importance of collocation relative to other components of word knowledge and their preference for a deductive or inductive approach to learning vocabulary.

### 6.3.1 Lexico-syntactic relations of words

The descriptive statistics in Table 6-2 demonstrate that the respondents inferred the meaning of an unknown word from contiguous words ( $M = 4.79$ ) as well as from the contextual information ( $M = 4.97$ ). They also reported that when learning a new word, they not only paid attention to the word itself but also its frequent collocates ( $M = 4.11$ ), grammatical functions ( $M = 4.56$ ) and the larger context in which it was embedded ( $M = 4.33$ ). Taken together, the respondents seemed to have taken into account lexico-syntactic relations as they learned new words, which was indicative of their awareness of the connections that a word had with the immediate and extended contexts. In other words, they reported having a certain degree of awareness that a word could not be seen as a discrete unit, but a node with lexico-grammatical, syntactic and discoursal associations (also see 2.4.1.1 for how connectionism conceptualizes the mental representation of word knowledge).

*Table 6-2 Learner awareness of lexico-syntactic relations of words*

Item	Mean	SD
12. When I encounter a new word, I infer its meaning from the immediately adjacent words.	4.79	0.80
13. When I encounter a new word, I infer its meaning from the textual context.	4.97	0.77
14. When learning a new word, I pay attention to the words that frequently co-occur with the new word. For example, <i>competition</i> : 'to face competition'; 'a fierce/intense competition'	4.11	0.82
15. When learning a new word, I like to know how it is used in a sentence. For example, <i>consist</i> : The category <u>consists of</u> 3 elements.	4.56	0.87
16. When learning a new word, I like to know how it is used in relation to the text.	4.33	0.80

Note: 6 - strongly agree; 5 - agree; 4 - toward agree; 3 - toward disagree; 2 - disagree; 1 - strongly disagree.

Nonetheless, while the respondents claimed having an awareness of the lexico-syntactic relations of words, their performance data has indicated otherwise. Section 4.1.1 has shown that before DALC intervention, the respondents scored 35.30% (Mean) on the collocation pretest in which the items tested the frequent collocations of the words they had studied previously, indicating a gap between their knowledge of collocations and that of form-meaning links. It appears that the respondents may not have paid as much attention to the lexico-syntactic relations of words as they claimed, or simply that they did not learn collocations well. Alternatively, such overestimated attention may be the result of the *halo effect*, because the participants may have tried to please the researcher by giving responses that they thought were desired (Mackey and Gass, 2005).



6.3.2 Perceived importance of components of word knowledge

Item 20 asked the respondents to rank six components of word knowledge in order, in terms of their perceived importance on a scale from 1 to 6, 1 being the most important and 6 being the least (see Appendix 13, Item 20). The mean rank assigned to each component is shown in the table below.

Table 6-3 Perceived importance of components of word knowledge

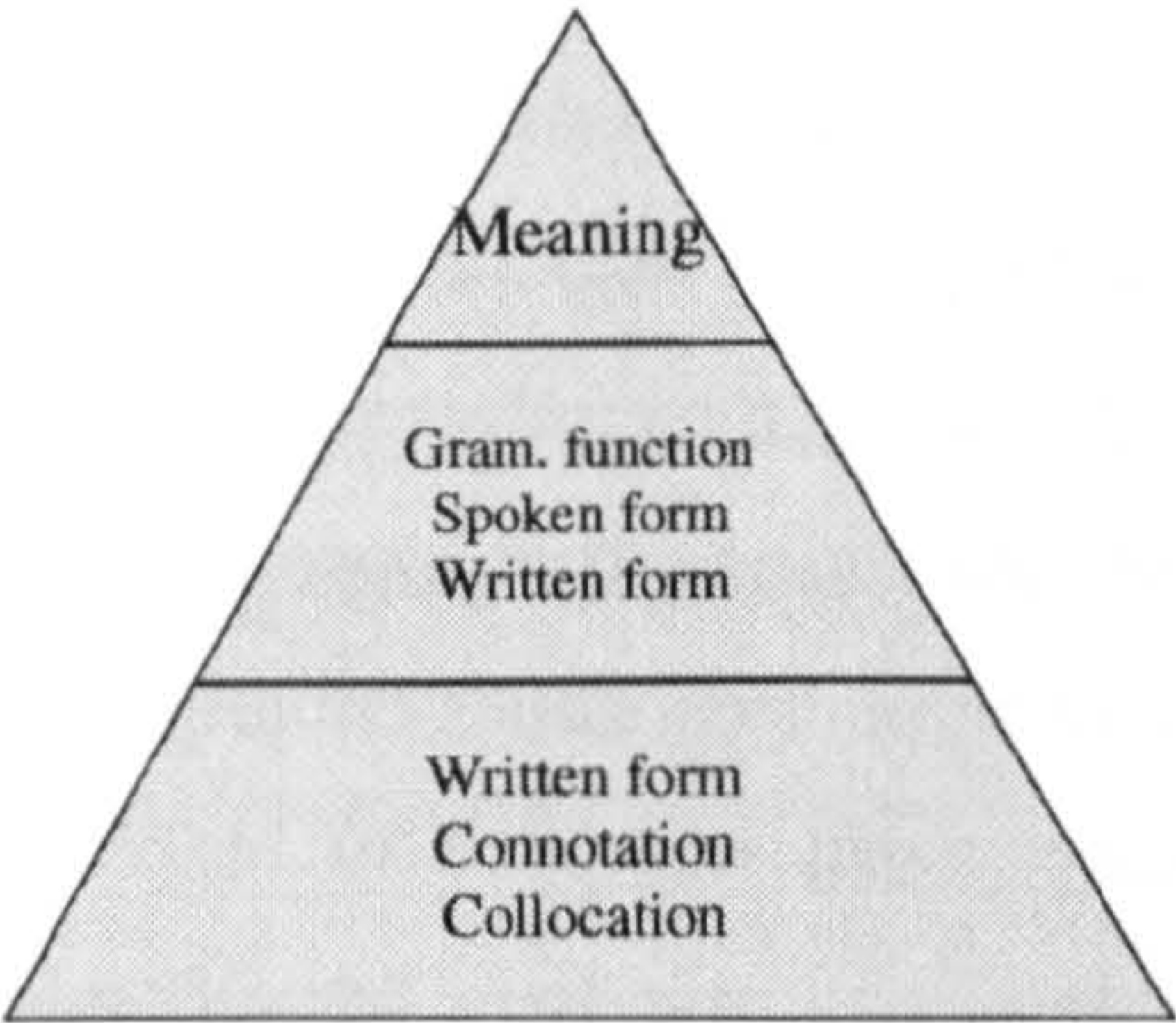
Component of word knowledge	Mean rank
Meaning	2.09
Grammatical function	3.21
Spoken form	3.27
Written form	3.86
Connotation	4.28
Collocation	4.29

Note: Rank 1 is the most important and rank 6 is the least important.

The Friedman’s ANOVA confirmed the existence of significant differences among the six components ( $p < .001$ , see Appendix 19). *Post hoc* Wilcoxon sign-ranked tests were then performed to further determine in which pairs of components the significant differences lay. In the *post hoc* tests for the Friedman’s ANOVA, the level of significance accepted was corrected as the critical level of significance divided by the number of comparisons: that is, 0.05 divided by the number of comparisons. In this case, 6 components constituted 15 pairs of comparisons, so the level of significance accepted here was 0.05 divided by 15, namely 0.0033. A Bonferroni correction was applied, so all effects were reported at a 0.0033 level of significance (Field, 2005). The inferential statistics above (see Appendix 19 for the results) reveal that the perceived importance of the components of word knowledge fell into three statistically distinct bands: Band 1 consisted of the most important component, *word meaning*, which was assigned a significantly higher rank ( $M = 2.09$ ) than other components; Band 2 was perceived to be of secondary importance, comprising *grammatical function* ( $M = 3.21$ ) and *spoken form* ( $M = 3.27$ ), which fell into the same band because they did not differ significantly from each other ( $p = .711$ ); perceived to assume only marginal importance, Band 3 was composed of *connotation* ( $M = 4.28$ ) and *collocation* ( $M = 4.29$ ). The two components did not differ significantly from each other ( $p = .624$ ), so they were both categorized into Band 3 in terms of perceived importance. Note that the component *written form* was assigned a mean rank of 3.86, falling in between Band 2 and 3, but did not differ significantly



from either band. Therefore, it was only sensible to classify it into both bands instead of either one: that is, it was perceived as only less important than the component *word meaning*, but as important as all the remaining components.



*Figure 6-1 The perceived hierarchy of components of word knowledge*

The perceived hierarchy of the components of word knowledge may account for the common misconception that a word is *learned* if the form-meaning link is known. Word meaning is perceived to be far more important than all the other components, because it is indisputably the most fundamental element of what constitutes word knowledge. The fact that the components *grammatical function*, *spoken form* and *written form* assumed secondary importance may be attributed to the pedagogical emphasis on grammatical and formal accuracy that was prevalent in language classrooms in Taiwan. Learners' word knowledge is often measured by their ability to spell a word correctly and recognize and/or recall the grammatical functions of a word. When it comes to assessment, a mis-spelled yet recognizable word is very likely to be less tolerable than a miscollocation or the misuse of a word in certain genres/registers. Consequently, deeper levels of word knowledge tend to be neglected, and this has an impact on the learners' awareness and learning of vocabulary. As highlighted in Section 2.2.1, in addition to form-meaning links, much more about words needs to be known. It is clear from the results that the importance of collocations has been underestimated, which calls for learner awareness-raising and adjustments to vocabulary pedagogy and assessment.



### 6.3.3 Deductive and inductive approaches to learning vocabulary

The statements in Items 17-19 referred to three distinct approaches to learning vocabulary (i.e., the deductive/top-down, inductive/bottom-up and integrated approaches), so it was assumed that the responses to these items were mutually exclusive to a greater or lesser degree, particularly Items 17 and 18 (see Table 6-4). That is, it was presupposed that if the respondent rated higher on Item 17, s/he was likely to rate lower on Item 18, as they referred to two opposite ways of learning vocabulary (i.e., a deductive and an inductive approaches). However, the results show that the respondents found all three approaches equally applicable to their vocabulary learning experience, yielding mean ratings 4.39, 4.27 and 4.59 respectively. The Wilcoxon signed-rank test indicates that Item 17 and Item 18 differed respectively from Item 19 in a statistically significant way (i.e. Item 17 - Item 19,  $p = .030$ ; Item 18 - Item 19,  $p < .001$ ), whereas they did not differ significantly from each other (Item 17 - Item 18,  $p = .312$ ). In other words, the informants claimed having similar degrees of preference for learning vocabulary *deductively* (as operationalized in Item 17) and *inductively* (Item 18). Nonetheless, the integrated approach (Item 19) was felt to be even more applicable to their experience than either approach alone. Respondents did not seem to use a deductive or inductive approach exclusively, but may use them alternately as the situation demanded. Therefore, they perceived that among the three approaches, the integrated approach best described their vocabulary learning experience.

Table 6-4 Preferred approaches to learning vocabulary

Item	Mean	SD
17. I learn how to use a word by memorizing its grammatical functions and applying them.	4.39	0.85
18. I learn how to use a word through reading sample sentences, observing their structures, and then generalizing the grammatical functions from the sentences.	4.27	0.91
19. I learn how to use a word by knowing its grammatical functions as well as observing sample sentences whereby I can verify the grammatical functions.	4.59	0.91

Note: 6 - strongly agree; 5 - agree; 4 - toward agree; 3 - toward disagree; 2 - disagree; 1 - strongly disagree.

In their study on learners' corpus consultation on VN collocations, Chan and Liou (2005) found that when learning collocations, their learners preferred both the *deductive* and *integrated approaches* to the *inductive* one which was mediated by concordancers. Their learners were less enthusiastic about inducing collocations from concordances, as they found it time-consuming and difficult to locate the target



collocations in concordances. Similar reservations about the efficiency of inducing collocations from corpus data were echoed by the informants in this study (as will be discussed at length in Section 6.5.1.4). The present study explored learners' preference for the three approaches to learning *vocabulary*, while Chan and Liou (ibid.) focused specifically on these approaches to learning *collocations*. Taken together, learners generally prefer an integrated approach to learning vocabulary, but when it comes to learning collocations, the inductive approach assisted by corpus consultation is less favourable. This may be attributable to their concern over the efficiency of such an approach, albeit a further probe to verify this inference is needed.

#### 6.4 Awareness of collocations

As regards the learners' awareness of collocations, the respondents claimed having known the notion of collocation prior to the study ( $M = 4.01$ , see the table below).

Table 6-5 Awareness of collocations

Item	Mean	SD
22. I have already known the concept of collocation prior to this study.	4.01	1.23
23. Learning collocations enhances the accuracy of word choices.	4.92	0.89
24. Learning collocations enhances the fluency of written or spoken production.	4.34	0.87
25. Learning collocations enhances the appropriateness of word choices according to register/genre.	4.76	0.87
29. I will pay attention to collocations in future English learning.	4.90	0.81

Note: 6 - strongly agree; 5 - agree; 4 - toward agree; 3 - toward disagree; 2 - disagree; 1 - strongly disagree.

Having received the pedagogical intervention on collocation learning, the respondents were asked about their perceptions of the importance of developing collocational knowledge. As noted in Section 2.3.1, collocational knowledge plays an important role in enhancing *accuracy, fluency and appropriateness* in language production. Resonating with the literature, the respondents held that the accuracy of word choices was attributable to collocational knowledge ( $M = 4.92$ ). As collocation is an integral part of word knowledge and in some cases the meaning of a word even comes from its collocates (i.e., *collocational prosody*, see 2.2.2.1), it follows that knowing more about the full range of collocations of a word contributes to a better grasp of word knowledge. Without a developed sense of collocational knowledge, learners would have to express intended meanings with longer and thus more error-prone expressions (Hill, 2000). It is widely recognized that the more ready-made chunks learners have at

their disposal, the less time and effort is needed for encoding and decoding language, thereby enhancing *fluency* in language production (as discussed in 2.3.1). This commonly held belief was also reflected in the learners' responses ( $M = 4.34$ ) after they came into intensive contact with collocations for a period of time. In terms of *appropriateness*, collocations have contextual preferences to a varying degree. A developed sense of collocational knowledge allows learners to produce appropriate utterances as the context or register demands (as discussed in 2.3.1). This claim was also echoed by the respondents: learning collocations enhanced the appropriateness of word choices according to register/genre ( $M = 4.76$ ). As Bonk (2001:116) puts it, '[t]he acquisition of appropriate collocations (e.g., *administer a test*) would appear to be an essential part of acquiring and demonstrating a competence in that speech community, since it reflects a deep knowledge of the common lexis of the field'.

Generally speaking, the importance of developing collocational knowledge claimed in the literature (i.e., enhancing *accuracy*, *fluency*, and *appropriateness* in language production) were supported by learner perceptions in this study. Given the importance attached to collocational knowledge, the learners expressed willingness to devote more attention to collocations in future English learning ( $M = 4.90$ ).

## **6.5 Perceptions of DALC**

This section presents evidence of the informants' experience and perceptions of exploiting corpus resources to aid collocation learning.

### **6.5.1 Hands-on experiences of DALC**

This subsection addresses the learners' hands-on experiences of corpus consultation, in terms of perceived level of difficulty, amount and authenticity of corpus data, and overall experience of DALC.

#### **6.5.1.1 Perceived level of difficulty of corpus data**

As shown in the table below, concordances did not seem to be too difficult to read for the respondents, as evidenced by the positive responses to Item 34 ( $M = 4.06$ ). The informants did not experience much difficulty understanding most of the

concordances ( $M = 4.12$ ). They were able to generalize target collocations, irrespective of incomplete concordancing lines ( $M = 4.37$ ). The common concern that the ‘messiness’ of corpus data would cause comprehension difficulties (see 2.3.2.2) was not reflected in the responses here.

*Table 6-6 Perceived level of difficulty of corpus data*

Item	Mean	SD
34. I find the difficulty level of concordancing output adequate.	4.06	0.70
35. I understand most of the concordancing output.	4.12	0.76
36. Although most concordancing output is displayed in incomplete sentences, it is sufficient to generalize frequent collocations of the target word.	4.37	0.88

Note: 6 - strongly agree; 5 - agree; 4 - toward agree; 3 - toward disagree; 2 - disagree; 1 - strongly disagree.

There has been a concern over the *decontextualization* of concordances for not being able to provide contextual information necessary for understanding the texts (Aston, 1995; Widdowson, 2000). As Aston (1995:260) argues, ‘in so far as it involves the decontextualization of individual instances from their original communicative setting, concordance analysis fails to engage the learner in those processes of negotiating meaning in its pragmatic aspects [...]’. Whereas it holds true that fragments of texts that are void of communicative contexts can be problematic when negotiation of meaning is involved, *decontextualization* did not seem to pose a major problem for the learners in this study where the purpose of consulting concordances was to generalize collocational patterns. Simply put, observing the immediate textual contexts may have sufficed to generalize collocational patterns. Cobb (1999) argues that the decontextualized concordance lines may in principle provide an opportunity to focus on contextual word learning (or collocation inducing in this case) without being distracted by the flow of discourse. It seems that the *breadth* of corpus data compensated for a lack of *depth*: that is, the profusion of corpus data made it possible to generalize salient patterns by browsing through many instances without having to understand in detail how each one was used in an extended context. In other words, even if one did not understand some of the concordance lines, s/he could turn to many more others and was still able to induce prominent patterns eventually.

In Yoon and Hirvela’s (2004) study on ESL learners’ attitudes toward corpus use (i.e., *Collins COBUILD Corpus*) in L2 writing, the participants were also reported to have no difficulty understanding the concordancing output. In contrast, Sun (2007) looked into EFL learners’ perceptions of a concordancing tool for academic writing using a



purpose-built, discipline-specific corpus, and found that some participants felt the language data in the corpus overwhelming. The perceived level of difficulty of corpus data depends as much on the texts in the corpus as on the attributes of the user, such as language proficiency, computer literacy or learning style. In this study, corpus data did not seem to cause comprehension difficulties for the respondents, it may well be that the language point (i.e., lexical collocation) was so specific that the learner only needed to observe the contiguous words without having to read or understand the extended contexts.

### 6.5.1.2 The amount of corpus data

As far as the pedagogical use of corpora is concerned, the *amount* of corpus data has been the subject of much debate. The respondents' perception on this issue also seems to be a mixed one. Generally, the respondents found the amount of concordances adequate, as indicated by a mean rating of 3.78 (see Table 6-7). The one sample t-test shows that the rating was significantly higher than the mid-point 3.5 ( $p = .001$ ), indicating a positive attitude (see 3.7.3 for the rationale for a cut-off point on the 6-point scale). However, a further probe with negatively phrased statements (Items 43 and 44) reveals that the respondents did not feel as positive about the amount of concordances as was indicated earlier at a general level. When asked if the amount of concordancing results was too large or too small, the respondents expressed a less optimistic attitude ( $M = 3.01$  and  $3.19$  respectively). The ratings were significantly lower than the mid-point 3.5 ( $p < .001$  and  $p = .002$  respectively), indicating a negative attitude. While these two items used conflicting statements, the respondents' equally negative attitude toward both statements was not necessarily self-contradictory. The participants were given six web-based concordancers to work with, so they might feel differently about the amount of corpus data in different concordancers. It is likely that they found some too large to manage and others too small to allow generalizations. Since the study did not set out to evaluate the concordancers *per se*, the questionnaire did not ask the respondents to comment on the corpus size of each concordancer. In addition, the frequency of occurrence of a collocation may also affect the ease of corpus search: some collocations are more frequent and therefore easier to induce from concordances than others. Taken together, the respondents may have encountered both situations where the corpus size affected

the ease of inducing collocations. Nonetheless, many of them (76.3%) felt that a larger corpus size would increase the success of corpus consultation.

*Table 6-7 Perceptions of the amount of corpus data*

Item	Mean	SD
33. I find the amount of concordancing output adequate.	3.78	0.82
43. The amount of search results is too large to manage.	3.01	0.95
44. The amount of search results is too small to generalize collocations.	3.19	0.95
49. What other feature(s) do you think a web-based concordancer should provide to better assist collocation learning? (Choose as many as applicable)		
user-friendly interface <input type="checkbox"/> <b>large corpus size <input checked="" type="checkbox"/> (76.3%)</b> bilingual search <input type="checkbox"/> results displayed in complete sentences <input type="checkbox"/>	collocation search <input type="checkbox"/> online dictionary <input type="checkbox"/> part-of-speech search <input type="checkbox"/> sorting function <input type="checkbox"/>	

Note: 6 - strongly agree; 5 - agree; 4 - toward agree; 3 - toward disagree; 2 - disagree; 1 - strongly disagree.

As far as corpus size was concerned, Chan and Liou's (2005) respondents had mixed feelings: approximately half of them (53.33%) felt that an increase in corpus size was necessary, while the other half did not. Their result contrasted with that in the present study where a fair proportion of respondents (76.30%) found a larger corpus size necessary in corpus consultation. Interestingly, the participants in this study were given six concordancers to work with, while those in Chan and Liou (ibid.) used only one of the concordancers. The respondents in this study had the opportunity to experiment with several concordancers, it is likely that they based their conclusion on the comparison and contrast across concordancers: search experiences with larger corpora may have been relatively easier or more successful, as compared with smaller ones, and thus led many of them to favour large corpora. Likewise, in Cheng, Warren and Xu's (2003) study on undergraduate ESL learners' direct use of corpora as a language study and language learning tool, participants were trained to use a range of mega-size corpora, including *Bank of English*, *Brown Corpus*, *LOB Corpus*, *British National Corpus (BNC)*, and the British English component of the *International Corpus of English (ICE-GB)*. Although the participants were given access to a number of large corpora, they still felt the need for even larger and more comprehensive corpora to better support corpus-based language study and language learning.

As demonstrated above, many learners perceive the need for a larger amount of corpus data after having a hands-on experience of corpus consultation. On the

contrary, scholarly discussion suggests the opposite regarding corpus size for pedagogical use. Aston (1995:54) argues that 'the virtues of large corpora seem less readily apparent' from a language learning perspective, suggesting a corpus size of 20,000 to 200,000 words is sufficient and preferable in learning contexts. Similarly, Braun (2005) maintains that a large corpus produces too many results, and the evaluation of the results can be time-consuming. Alternatives to counter the problems posed by corpus size include the use of sub-corpora within large corpora (cf. Aston, 1995), the use of small genre-specific corpora (cf. Ghadessy, 1989) or purpose-built corpora compiled by learners and/or teachers themselves (cf. Aston, 1995; Tribble, 1997). Purpose-built corpora, referred to as 'quick-and-dirty solutions' by Tribble (1997), may not be able to meet the standards of professional corpus creation, but may be more conducive to learning 'insofar as they can be specifically targeted to the learner's knowledge and concern' (Aston, 1995:9). Gavioli and Aston (2001) suggest teachers to grade corpus-assisted language learning activities, starting from smaller corpora with less varied text types and moving on to larger and more heterogeneous corpora.

### 6.5.1.3 Perceived authenticity

A review of the literature shows that the major strength afforded by corpora to language learning is the real language data used for genuine communication purposes, as opposed to that constructed for pedagogical purposes (see 2.3.2.1 for discussion on *genuineness vs. authenticity*). In this study, the learners' positive attitude toward the language data in corpora lends support to the assertion above. The table below shows that the respondents perceived corpus language to be more authentic than textbook language, and that they had more confidence in the former, as indicated by the positive responses to Items 31 and 32 ( $M = 4.24$  and  $4.15$  respectively).

**Table 6-8** *Perceptions of corpus language*

Item	Mean	SD
31. I think the language in corpora is more authentic than that in coursebooks.	4.24	0.77
32. I think the language in corpora is more credible than that in coursebooks.	4.15	0.77

Note: 6 - strongly agree; 5 - agree; 4 - toward agree; 3 - toward disagree; 2 - disagree; 1 - strongly disagree.

L2 textbooks have long been criticized for a poor representation of the language use in real-life contexts as a result of pedagogic adaptation or construction (Gilmore,



2007). The respondents' perception here was attested by a body of empirical evidence demonstrating mismatches between various aspects of real language use (mostly sampled from large corpora) and that in textbooks (cf. Tsui, 2004). The result here shows that when given the opportunity to be acquainted with corpus language, the respondents seemed to find it more truthfully representative of real language use than the textbook language is.

Few studies have explored learners' opinions on the authenticity of corpus language, with the exception of Chambers (2005) and Farr (2008). Chambers (ibid.) examined 14 language learners' corpus consultation processes, and found a general consensus as to the perceived authenticity and richness of corpus language, a deeper processing of language as a result of discovery learning, and learner empowerment (learner autonomy is addressed in 6.5.2.2). Some of her participants described corpus language as 'authentic', 'real', 'up-to-date' and 'relevant', as opposed to the invented examples in coursebooks. They also appreciated the large number of examples provided in corpora, in contrast to the limited number in coursebooks. Notwithstanding the favourable attitude toward real language in corpora, her participants did not see corpora as a replacement for textbooks. Along the same lines, Farr (ibid.) found that student teachers perceived the access to real language data as the major strength of corpus-assisted language learning, among others such as promoting the spirit of inquiry and raising language awareness. Acknowledging the limitations of both corpus language and textbook language, Johns (1986:159) maintains that corpus language that is systematically organized by concordancers provides a balance between 'the highly-organized, graded, and idealized language of the typical coursebook, and the potentially confusing but far richer and more revealing "full flood" of authentic communication'. As noted in Section 2.4.3, EFL contexts may be limited in providing learners with naturalistic exposure to English, and the language in the classroom and/or textbooks often does not, and is not perceived to reflect how it is used in real-life situations, so corpora can offer rich resources of language in use provided that adequate pedagogical mediation is given.

As highlighted in Section 2.3.2.1, *genuineness* is an inherent quality of corpus data, whereas *authenticity* needs to be built up through perceived relevance to the learner. Pedagogical mediation is essential to contextually reconstitute corpus findings to

create relevance to the learner. In the present study, learners perceived corpus language to be *authentic*, probably because *authenticity* was taken to be an inherent quality: the questionnaire was written in the respondents' L1, Chinese, which did not distinguish between *authenticity* and *genuineness*. Alternatively, it may well be that DALC intervention created relevance to the learner: the target collocations were based on the words studied previously, thereby expanding learners' prior word knowledge; also, the tasks were specifically designed for this group of participants, customizing the content according to their levels of language proficiency and learning needs. Such pedagogical mediation may have contributed to learners' positive attitude toward the authenticity of corpus language.

#### 6.5.1.4 The overall perception of DALC

As shown in Table 6-9, the informants found the two major mediating tools of DALC helpful in facilitating collocation learning: the collocation worksheets ( $M = 4.88$ ) and the web-based concordancers ( $M = 4.96$ ). However, when asked about the perceived level of difficulty in generalizing collocations from concordances, the informants expressed a neutral attitude ( $M = 3.43$ ). The one sample t-test reveals that the mean rating 3.43 did not differ significantly from the midpoint 3.5 on a 6-point scale ( $p = .393$ ), suggesting a genuinely neutral attitude. Nonetheless, they held reservations about the time needed for such endeavours ( $M = 3.20$ ). The rating was significantly lower than the mid-point 3.5 ( $p = .005$ ), indicating a negative attitude.

Table 6-9 Overall perceptions of DALC

Item	Mean	SD
26. The concordancers are helpful in learning collocations.	4.96	0.87
27. The collocation learning worksheets are helpful in learning collocations.	4.88	0.85
41. It is difficult to generalize collocations from corpus data.	3.43	0.83
42. It is time-consuming to generalize collocations from corpus data.	3.20	1.04
28. This learning experience raises my awareness of collocation.	4.81	0.83
30. I will keep using web-based concordancers to assist my collocation learning.	4.77	1.00

Note: 6 - strongly agree; 5 - agree; 4 - toward agree; 3 - toward disagree; 2 - disagree; 1 - strongly disagree.

The concern over the efficiency of corpus consultation was echoed in Chambers (2005): her participants described such processes as 'tedious', 'time-consuming' or 'laborious'. Admittedly, the induction processes are by nature painstaking and laborious, so the learners' concern over efficiency is to be expected. However, it is precisely the time and effort devoted into the induction processes that leads to deeper

processing of language input, thereby increasing the likelihood of converting *input* into *intake*. As discussed in Section 2.4.1.2, empirical evidence found in the field of cognitive psychology demonstrates that cognitively demanding activities which require a more intense manipulation of information yield more effective learning (Craik and Lockhart, 1972; Craik and Tulving, 1975). Despite the informants' concern over the time and effort needed to induce collocations from corpora, they made considerable progress at all levels of collocational knowledge, as evidenced in Chapter 4.

Broadly speaking, notwithstanding the concern over corpus size (see 6.5.1.2) and the efficiency of corpus consultation, the respondents recognized the benefits of such an endeavour on raising an awareness of collocations ( $M = 4.81$ ). Most participants were also optimistic about the prospect of continual use of concordancers to assist collocation learning ( $M = 4.77$ ).

## **6.5.2 Perceived benefits of pedagogical use of corpora**

The following subsections discuss the learners' perceived value of pedagogical use of corpora in vocabulary learning (6.5.2.1) and language learning in general (6.5.2.2).

### **6.5.2.1 Vocabulary learning**

The informants generally held a positive attitude toward corpus use in assisting vocabulary learning, and recognized its value in *consolidating* and *enhancing* word knowledge (see Table 6-10). Through the alignment of concordances, corpora are able to provide learners with intensive exposure to multiple instances of a query word, thereby helping to *consolidate* knowledge of word use. The respondents agreed that the wealth of language data in corpora was conducive to *consolidating* knowledge of the words studied previously ( $M = 4.31$ ). Schmitt (2008) stresses that 'recycling' previously studied words in a principled way for consolidation is important in vocabulary learning. The profusion of corpus data promises to provide rich resources for recycling, be it through pedagogical mediation or self-access consultation. More specifically, the informants also accepted that corpus consultation not only *consolidated* but also *enhanced* knowledge of the words studied previously ( $M = 4.58$ ). Regarding the ways in which corpus consultation enhanced word knowledge,



the respondents felt that they came into contact with a broader range of collocates (M = 4.64) of the target words, and was sensitized to the nuances of near-synonyms (M = 4.20) by observing many instances of the same word.

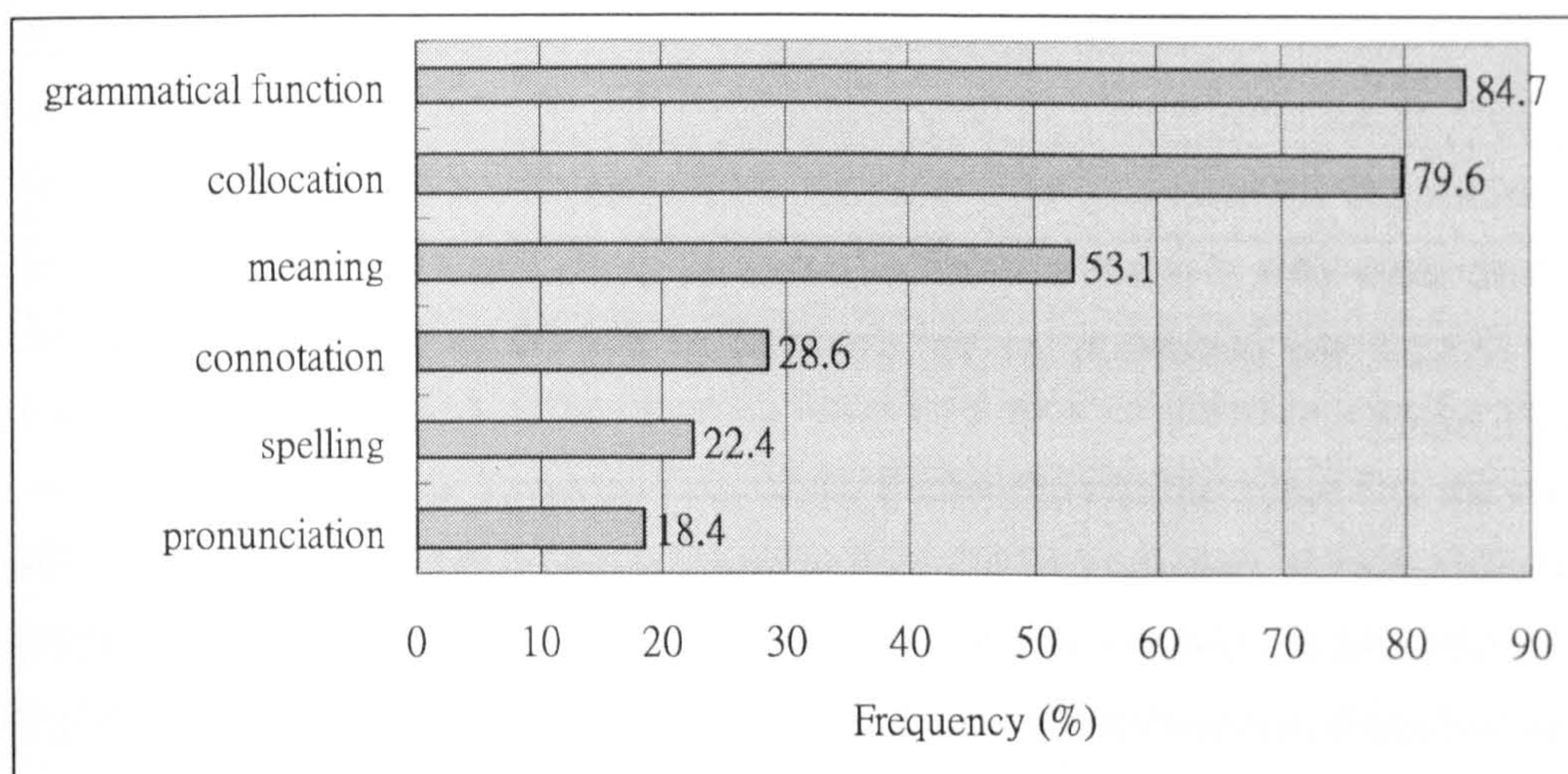
**Table 6-10    Perceived benefits of corpus consultation on vocabulary learning**

Item	Mean	SD
40. Corpus consultation consolidates the knowledge of the words studied previously.	4.31	0.95
46. Corpus consultation enhances the knowledge of the words studied previously.	4.58	0.79
37. Observing concordancing output allows me to discover a range of collocates of the target word.	4.64	0.74
38. Observing concordancing output helps me to notice the nuances between near-synonyms.	4.20	0.80

Note: 6 - strongly agree; 5 - agree; 4 - toward agree; 3 - toward disagree; 2 - disagree; 1 - strongly disagree.

The results above have shown that the respondents recognized concordancing as beneficial in *consolidating* and *enhancing* word knowledge. What follows demonstrates how concordancing was perceived as to its usefulness in consulting various aspects of word knowledge, including *written form*, *spoken form*, *meaning*, *connotation*, *collocation* and *grammatical function*. The Cochran’s test was performed to examine the differences in the perceived usefulness of concordancing for the six components above. The result shows significant differences among these six components ( $p < .001$ , see Appendix 19). To follow up, a *post hoc* procedure, the McNemar’s test, was performed to make multiple pair-wise comparisons, thereby determining the differences between the components in pairs (see the results in Appendix 19). A Bonferroni correction was applied, so all effects were reported at a 0.0033 level of significance (0.05 divided by 15 pair-wise comparisons equalled 0.0033). See Section 6.3.2 for the rationale for the Bonferroni correction.





**Figure 6-2** *Perceived usefulness of corpus consultation for components of word knowledge*

As illustrated in the figure above, the degrees of perceived usefulness of concordancing for enhancing knowledge of grammatical functions and collocations were both very high (84.7% and 79.6% respectively), and the degrees did not differ statistically from each other ( $p = .359$ ). That is, concordancing was perceived to be highly useful for consulting the grammatical functions and collocations of a word. As far as word meaning was concerned, the responses were divided, with only half of the informants (53.1%) thinking word meanings could be accessed through corpus consultation. Even though the respondents seemed to hold reservations about the usefulness of concordancers as a resource for consulting word meanings, its degree of perceived usefulness was still significantly higher than those of the remaining components, namely connotation, written and spoken form (see Appendix 19 for the McNemar's test results). Concordancers were regarded as having only limited usefulness for facilitating the learning of connotations, written and spoken form of a word (28.6%, 22.4% and 18.4% respectively). The degrees of perceived usefulness of concordancers in facilitating knowledge of these three components did not differ significantly from one another (see Appendix 19).

Overall, the degrees of perceived usefulness of concordancers in facilitating various aspects of word knowledge fell into three levels: highly useful, useful and less useful. Concordancers were perceived to be a powerful reference tool for accessing information on grammatical functions and collocations, which fell neatly into the broad category of *word use* within the word knowledge framework in Nation (2001) (see Table 2-1), as distinct from the categories of *word meaning* (form-meaning link,



connotation) and *word form* (spoken and written). The fact that the respondents found concordancing most helpful in enhancing knowledge of *word use* (e.g., grammatical functions and collocations) coincided with the presupposition that concordances were able to show how a word was typically used by providing a multitude of instances (see 2.3.2.2 for discussion on the potential of corpora in highlighting linguistic regularities).

Opinions divided as to the usefulness of concordancing for accessing word meanings. It may be that, compared with corpora, dictionaries seem to be a more straightforward source for accessing information on *form-meaning links*, albeit rather limited in providing examples of *word use*.

Whereas the core meaning of a word can easily be found in dictionaries, the nuances in connotations need to be illuminated by the contexts in which the word is actually used. In this sense, concordances can be usefully exploited to illustrate the connotations of a word. However, the informants did not think so, as they rated concordancing as being of limited use in enhancing knowledge of connotations. Such a low degree of perceived usefulness may be attributed to the task demands in this study: less marked aspects of word knowledge, such as connotations, might have been eclipsed by the focus on collocation learning. Also, learners did not seem to find concordancing helpful in enhancing knowledge of *word form*, which is reasonable given that corpus consultation would not have been possible if the written form was unknown to the user in the first place.

#### **6.5.2.2 Language learning**

In addition to providing language data for *induction*, corpora can be used for *deduction*, namely verifying the linguistic hypothesis the user has in mind. The informants generally agreed that corpora were useful for verifying such linguistic hypotheses ( $M = 4.60$ ). As concordances are extracted from texts used for real communication, learner may compare what they have in the interlanguage with what is actually used in the TL using corpus data for verification. Consistent with the current finding, Yoon and Hirvela (2004) also found that their participants felt confident using corpora as a tool for testing linguistic hypothesis.



**Table 6-11 Perceived benefits of corpus consultation on language learning**

Item	Mean	SD
39. Concordancers help me verify my linguistic hypotheses.	4.60	0.84
45. Generalizing collocations from corpus data enhances my sense of autonomous learning.	4.42	0.80

Note: 6 - strongly agree; 5 - agree; 4 - toward agree; 3 - toward disagree; 2 - disagree; 1 - strongly disagree.

It is generally accepted that learners' self-access corpus exploration contributes to *autonomous learning* (as discussed in 2.3.2.3). Bernardini (2001) argues that the skills and strategies that can be developed through corpus consultation are conducive to independent and autonomous learning. The actual extent to which corpus consultation skills can be transferred to contribute to autonomous learning in general has not been empirically verified, but the concern here is how learners themselves perceive such endeavours in contributing to learner autonomy. The result reveals that the induction processes aided by corpora were positively perceived to enhance a sense of learner autonomy ( $M = 4.42$ ). As noted in Section 6.5.1.3, Chambers' (2005) participants were reported to feel a sense of learner empowerment through corpus consultation. Learner empowerment perceived by Chambers' (ibid.) participants to some extent coincided with the sense of learner autonomy discussed here. Likewise, O'Sullivan (2007) cited learners' comments such as '[a] fairly reliable source for independent learning' to demonstrate their recognition of corpus consultation as beneficial for autonomous and independent learning. Bernardini (2002) and Yoon and Hirvela (2004) also found learner perceptions of increased confidence arising from self-access corpus use (see 2.3.2.3). Taken together, the above positive attitudes, be it learner empowerment or increased confidence in learning, give an indication that corpus consultation goes a long way toward autonomous learning.

### 6.5.3 Perceptions of web-based concordancers

The informants did not report having much difficulty in using web-based concordancers ( $M = 4.05$ , see the table below). They also found the interfaces of the concordancers user-friendly ( $M = 3.84$ ). In the training session, the learners were given a handout introducing the functions of the concordancers, so using these tools did not seem to cause difficulties for them (see Appendix 4 for the handout).

**Table 6-12 Perceived accessibility of web-based concordancers**

Item	Mean	SD
47. The interfaces of the web-based concordancers are user-friendly.	3.84	0.92
48. The web-based concordancing tools are easy to use.	4.05	0.90

Note: 6 - strongly agree; 5 - agree; 4 - toward agree; 3 - toward disagree; 2 - disagree; 1 - strongly disagree.

The following subsections present learner perceptions of important concordancer features in collocation induction, and their preferred concordancers.

### 6.5.3.1 Concordancer features

Item 49 asked the respondents to choose the concordancer features that they found important in inducing collocational patterns. The features given are as follows:

- A user-friendly interface;
- A large corpus size;
- A bilingual (Chinese/English) search function;
- Concordances displayed in complete sentences;
- A collocation search function;
- A built-in dictionary;
- The part-of-speech (POS) search function;
- Alphabetical sorting by contiguous words.

As shown in Table 6-13, a user-friendly interface (89.7%) was perceived to be far more important than other features, followed by a large corpus size and the POS search function, both of which were perceived as important by 76.3% of the respondents. While large corpora provide more language data for generalization, the POS search function allows users to narrow the scope of search by specifying the word class of the query word and/or that of its collocate. 70.1% of the informants felt that a concordancer with a built-in dictionary would be helpful when searching for collocations in concordances. Whereas central to corpus consultation was inducing linguistic patterns from the bottom up, 69.1% of the respondents would like target collocations to be generated by concordancers automatically. A bilingual search function in concordancers was perceived necessary by 66% of the respondents. Over half of the respondents felt that concordances displayed in complete sentences and an alphabetical sorting function would be helpful (55.7% and 53.6% respectively). It was surprising that the former was attached relatively less importance than other functions, considering the scholarly criticism about the incompleteness and decontextualization of concordances (see 6.5.1.1). This result along with that of Item 36 (see 6.5.1.1) seems to suggest that the learner perception did not agree with the scholarly concern

over the decontextualized corpus language insofar as collocation induction was concerned: generalizing collocational patterns may not require the comprehension of concordances at the syntactic and/or discoursal levels.

**Table 6-13** *Perceived importance of concordancer features*

Interface	Larger corpus	POS search	Online dictionary	Collocation search	Bilingual	Complete sentences	Alphabetical sorting
89.7%	76.3%	76.3%	70.1%	69.1%	66.0%	55.7%	53.6%

With regard to the possible improvements that can be made to concordancers, Chan and Liou (2005) found similar responses from their participants. The vast majority of the respondents (93.3%) expressed the need for a collocation search function which automatically generated collocates of a query word, instead of concordances from which collocations had to be induced by the user. Most of their respondents (76.7%) also found the POS search function necessary to increase the efficiency of concordancing. As illustrated above, the results of the present study also show similar responses as to the necessity of a collocation search function and POS search function in assisting corpus consultation (69.1% and 76.3% respectively). The finding is not surprising, as these two features narrow the search scope and reduce search time, thus enhancing the efficiency of corpus consultation. The POS search function is highly favourable, because it narrows the search scope by specifying the word class of the query word and/or that of the collocate so that the user would not have to go through all concordance lines to locate the target combinations. With the function of collocation search, a concordancer automatically generates collocates of the query word, saving the user the time and effort of manually inducing collocations. The respondents' favourable attitude toward such functions may have resulted from their concern about the efficiency of manually generalizing collocations from corpus data (see 6.5.1.4).

### 6.5.3.2 User preferences for concordancers

As there were six web-based concordancers designated for DALC intervention sessions, Item 50 asked the respondents to choose the one(s) that they favoured. The results are shown in the figure below. The Cochran's test and *post-hoc* McNemar's test reveal significant differences among the degrees of preference for the six concordancers (see Appendix 19 for the inferential statistics). A Bonferroni correction



was applied, so all effects were reported at a 0.0033 level of significance (0.05 divided by 15 pair-wise comparisons equalled 0.0033). See Section 6.3.2 for the rationale for the Bonferroni correction.

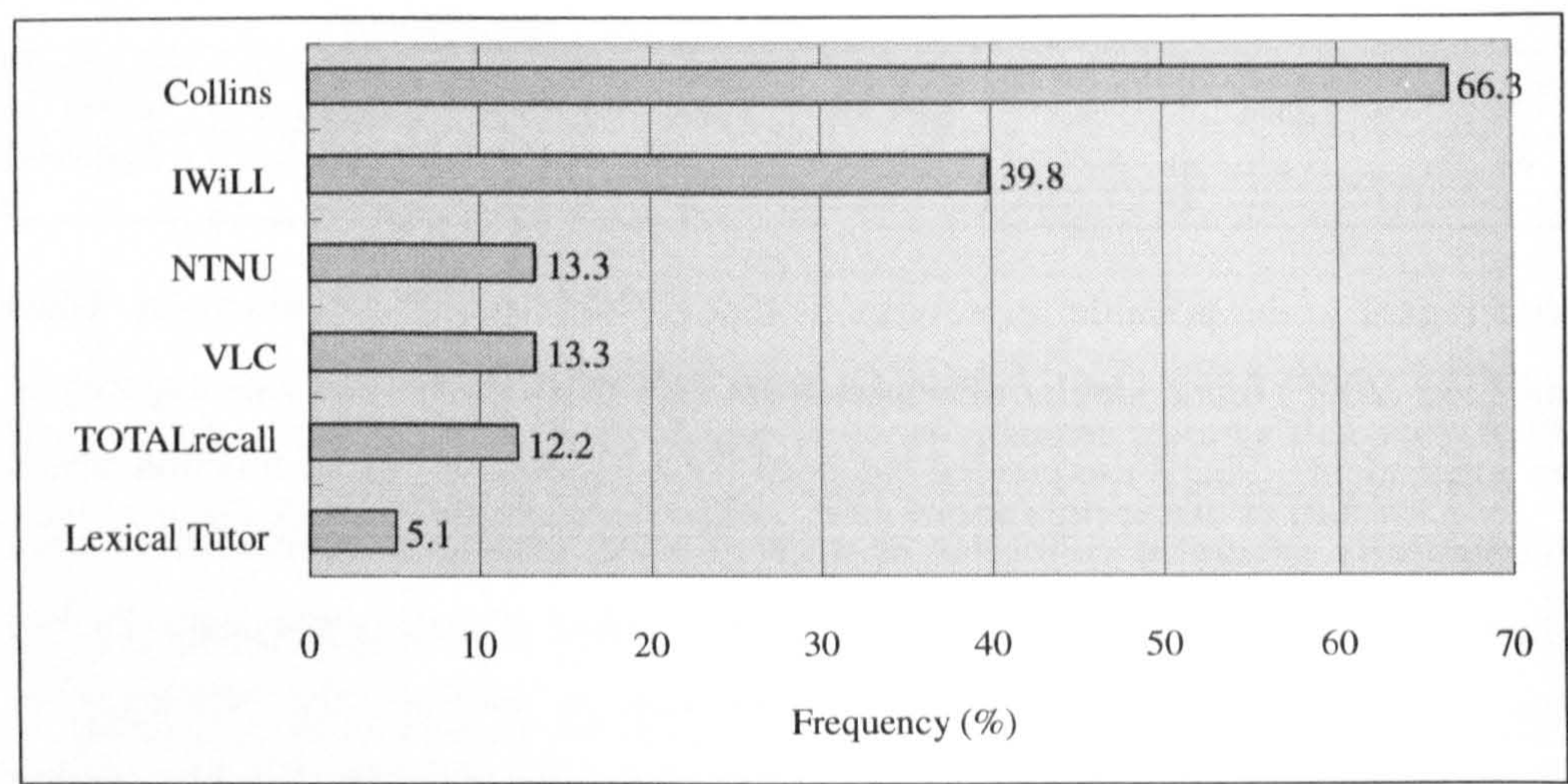


Figure 6-3 Degrees of preference for the designated concordancers

*Collins WordbanksOnline* topped the list of concordancers, because over half of the respondents (66.3%) chose this tool as the most preferred one. Among the respondents who chose this one, many indicated that it was user-friendly. *IWiLL* was favoured by 39.8% of the respondents, a proportion significantly lower than that of *Collins WordbanksOnline* ( $p = .001$ , see Appendix 19) but higher than the remaining four concordancers (see Appendix 19 for the four pair-wise comparisons,  $p < .001$ ). The inferential statistics demonstrate that *IWiLL* was considered as the second most favoured concordancer, albeit not as preferable as *Collins WordbanksOnline*. The remaining four concordancers, *NTNU Web Concordancer*, *VLC Web Concordancer*, *TOTALrecall* and *Lexical Tutor*, though accounting for varying degrees of preference (13.3%, 13.3%, 12.2% and 5.1% respectively), did not differ significantly in a statistical sense.

*Collins WordbanksOnline* was by far the most favoured concordancing tool because of its clarity and user-friendliness as suggested, followed by *IWiLL*, whose major strength lay in its Chinese/English bilingual interface and POS search function. Both concordancers provided POS search function which narrowed the scope of corpus search. The degrees of preference for the remaining four concordancers were



surprisingly low, as both *NTNU Web concordancer* and *TOTALrecall* had Chinese/English bilingual interfaces, and the latter even provided concordances in both languages. It seems that the respondents relied heavily on POS search function when concordancing, and that parallel L1 concordances did not appeal to this particular group of learners.

## 6.6 Summary

This chapter has addressed the issues on how learners responded to DALC. Most respondents exploited computer and online resources for English learning with a varying degree, but relatively few used web-based concordancers. Having been acquainted with web-based concordancers, the participants generally recognized the merits of corpus use in assisting language learning: the major strengths of pedagogical use of corpora in L2 contexts lay in the *authenticity* and *profusion* of language data, as opposed to the limited, invented examples in textbooks. Learners found corpus consultation helpful for *consolidating* and *enhancing* word knowledge. More specifically, corpus consultation was perceived to be particularly useful in providing information on *word use* (e.g., grammatical functions and collocations), as opposed to that of *word meaning* and *word form*. In addition to the linguistic level, corpus consultation was also perceived to foster a sense of learner autonomy. Hence, the respondents felt confident about the continual use of concordancers in assisting their language learning, with increased attention devoted to collocations.

With regard to collocational awareness, learners were sensitized to the importance of a developed sense of collocational knowledge: that is, it contributed to *accuracy*, *fluency* and *appropriateness* in language production. They also reflected that they paid attention to the lexico-syntactic relations when learning new words, and that the combination of the deductive and inductive approaches was most applicable to their vocabulary learning experience.

In terms of the hands-on experience of corpus consultation, the learners found the concordancers easily accessible, and they reported having a good command of concordancing skills. The amount, format and difficulty level of concordances did not seem to pose difficulties for the users. They favoured two concordancers in particular, *Collins WordbanksOnline* and *IWiLL*, because they had POS search function which



efficiently narrowed the scope of corpus search. Notwithstanding the reported ease of corpus consultation and the perceived benefits, the respondents were less optimistic about the time and effort required for such an endeavour.

Generally speaking, the results demonstrate an overall positive attitude toward the data-driven approach to learning collocations and/or language in general. Given the theoretical and empirical evidence, corpora hold considerable promise for assisting language learning. However, there was a discrepancy between the perceived benefits of corpus consultation and its efficiency. While the perceived efficiency can influence the willingness of using such reference tools, it has pedagogical implications for more systematic training on concordancing skills, or 'corpus consultation literacy' (O'Sullivan, 2007), in order to guide learners to make efficient use of corpus resources.



## Chapter 7 Conclusion

This chapter recapitulates on the key findings of the current research in relation to the three research questions (7.1). The limitations of this study are discussed, and directions for future research are indicated (7.2). The implications for FL pedagogy are also considered in Section 7.3. This chapter concludes with a brief summary of the research (7.4).

### 7.1 Key findings and discussion

In order to understand how corpus resources can be exploited to facilitate collocation learning in FL contexts, the present study examined a corpus-assisted pedagogical approach, namely DALC, from three dimensions: *learning product*, *learning processes* and *learner perceptions*.

#### 7.1.1 Learning product of DALC

RQ1: Does a data-driven approach to learning collocations facilitate EFL learners' development of collocational knowledge? If so, how does it facilitate such development?

The first key aim of this study is to examine the *learning product* or *learning outcomes* of DALC. The *learning product*, namely the changes that occurred in learners' collocational knowledge as a result of DALC intervention, was measured at three levels: *receptive*, *controlled productive* and *free productive knowledge*. Measurements were taken from both the experimental group and the control group before and after the intervention (or non-intervention) period (i.e., 2 groups x 2 points in time). Comparisons were made within and between groups, to assess the effectiveness of DALC on learners' collocational knowledge.

The results show that the two groups did not differ significantly in the baseline *receptive* and *controlled productive collocational knowledge*, but the experimental group made significant progress in these respects after receiving the intervention, whereas their control counterparts had remained at the same entry level throughout the period of time (see 4.1.1 and 4.1.2). The results are hardly surprising, considering that one group was given the intervention, while the other was not. However, the

extent of the progress made by the experimental participants (Mdn 43.43%→68.18% in *receptive knowledge*; Mdn 25.93%→57.14% in *controlled productive knowledge*, see 4.1.3) was tantalizing, lending support to the efficacy of DALC in terms of increasing EFL learners' *receptive* and *controlled productive knowledge* of lexical collocations.

With respect to *free productive collocational knowledge*, comparisons were made only within groups, and not between groups, as the differences in genre and topic of writing assignments rendered the two groups incomparable (as discussed in 3.5.2.2). Within-group comparisons indicate that the experimental group made significant increases in the *number* and *acceptability* of the collocations produced, whereas their control counterparts used approximately same number of collocations over time, but the acceptability thereof also increased significantly (see 4.2.1 and 4.2.3). In terms of the *types* of the collocations produced (see 4.2.2), the experimental group used fewer *free combinations* and more *restricted collocations* after the intervention, suggesting a heightened awareness of collocational restrictedness, or collocability; their non-intervention peers showed the opposite trend, by using even more *free combinations* and fewer *restricted collocations*. It is noteworthy that the control group's increase in *collocational acceptability rate* was not so much a sign of progress in collocational knowledge, but the result of an even larger proportion of *free combinations* used in writings, as *free combinations* were less restricted and thus less error-prone. To probe further into the *miscollocations* produced by the experimental group (see 4.2.4), not only did they produce fewer miscollocations after the intervention, but they also used fewer miscollocations attributable to L1 influences, implying an increased awareness of the language-specific nature of collocations.

Overall, DALC has been attested to develop, *quantitatively* and *qualitatively*, learners' *receptive*, *controlled productive* and *free productive knowledge* of lexical collocations. Whereas previous studies (e.g., Chan and Liou, 2005; Sun and Wang, 2003) only examined one or two levels of collocational knowledge, this study has gone a step further to demonstrate the outcomes of corpus-assisted collocation learning in a more comprehensive manner, by looking into learners' *free production* of collocations in addition to *recognition* and *cued recall*. However, it is worth noting that the results here show that corpus consultation extended all levels of collocational knowledge, but

does not go so far as to claim that it pushed collocational knowledge forward from the passive end to more active one (e.g., moving from *cued recall* to *free production*) on the receptive-productive continuum, because none of the target collocations tested for *recognition* and *controlled production* were used in the *free production*. The practical constraints on sampling learners' written production (as discussed in 3.5.2.2) precluded a further probe into how these target collocations would have been used in free production. While the present findings have established that DALC extends three levels of collocational knowledge respectively, it merits further research to investigate the ways in which corpus consultation advances collocational knowledge along the receptive-productive continuum.

As noted in Chapter 2 (2.4), DALC is grounded in cognitive approaches to SLA as well as in SCT, both paradigms contribute to account for how *learning* occurs with DALC. In terms of cognitive approaches to SLA, the *quantitative* and *qualitative* changes found in learners' collocational knowledge substantiate that DALC is able to *restructure* collocational knowledge. Gass and Selinker (2008:230) characterize *restructuring* as:

the changes made to internalized representations as a result of new learning. Changes that reflect restructuring are discontinuous or qualitatively different from a previous stage. Learning means the inclusion of additional information which must be organized and structured. Integrating new information into one's developing second language system necessitates changes to parts of the existing system, thereby restructuring, or reorganizing, the current system and creating a (slightly) new second language system. Mere addition of new elements does not constitute restructuring.

Judging by the characteristics of *restructuring*, DALC did bring about *quantitative* and *qualitative* changes in learner's collocational knowledge: *quantitative* changes were found in learner's *receptive*, *controlled productive* and *free productive collocational knowledge*; *qualitative* changes were evident in the increased awareness of *collocability* – learners were sensitized to the *usage-based* and *language-specific* nature of collocations. In view of such changes, DALC is attested to restructure collocational knowledge, thus facilitating the *learning* of collocations. This finding also lends support to SCT's claim that *symbolic mediation* may effect *quantitative* and *qualitative* changes in mental functions such as *learning* (see *symbolic mediation* in 2.4.2.2): the *symbolic mediation* arising from observing and analyzing real language in corpora did lead to increases in learners' knowledge and awareness of collocations. Underpinned by cognitive approaches to SLA and SCT, DALC is conducive to



developing collocational knowledge as well as raising collocational awareness.

### 7.1.2 Learning processes of DALC

RQ2. What is the nature of the thinking processes EFL learners engage in during the data-driven approach to learning collocations?

The performance data above demonstrates that corpus consultation *restructures* collocational knowledge by bringing about *quantitative* and *qualitative* changes. While the theoretical and empirical literature has well documented the effectiveness of corpus consultation on various aspects of language learning, there seems to be a missing link between *corpus consultation* and *language learning*: what transforms concordances on a computer screen to collocational connections within the learner's mental lexicon? The second key aim of this study is thus to uncover the *thinking processes* underlying the observable corpus consultation behaviour in order to understand what contributes to the *restructuring* of collocational knowledge. To achieve this aim, *mentalist data* (i.e., concurrent think-aloud verbal protocols) and *behaviouristic data* (i.e., corpus queries) were gathered to reconstruct thinking processes during corpus consultation.

The results of the behaviouristic dataset show that in addition to the six designated web-based concordancers, the participants also used a range of online resources to complement corpus search, including online dictionaries, translation tools, search engines and discussion boards (see 5.1.2). When consulting corpora, participants generally employed the following query methods: *node search*, *collocate search*, *POS search* and *collocation search* (see 5.1.1 for details). To approach the DALC task, learners consciously or unconsciously activated mental capacities to process corpus language input in working memory. A vast array of strategies was found to be used for such information processing: 1) *cognitive strategies* that directly manipulated language data (see 5.2) – *deduction*, *induction* (grouping and differentiating, inferencing from various knowledge sources), *substitution* (with de-lexicalized words or synonyms) and *translation*; and 2) *metacognitive strategies* that orchestrated the cognitive ones noted above (see 5.3) – *planning*, *monitoring* and *evaluation*. In particular, learners drew on a wide range of knowledge sources for *inferencing*: *contextual information*, *linguistic knowledge* (lexical and grammatical knowledge)

and *world/schematic knowledge*.

Corpus consultation endeavours were found to be highly idiosyncratic and discursive in nature, depending on various external factors (e.g., task demands, reference tools, search results) as well as internal ones (e.g., individual differences, prior knowledge), so they could not be easily reduced to any linear, standard model. In view of the idiosyncratic nature of corpus consultation, this investigation was exploratory and the discussion was descriptive with the aim of portraying such endeavours as truthfully as possible. In contrast, Kennedy and Miceli (2001) and Sun (2003) also looked into corpus consultation processes, but they both generalized linear steps of such processes (see 2.5.2 for both studies). As neither of these two studies combined *mentalistic* and *behaviouristic* datasets, as this study did, to uncover the highly complex corpus consultation processes, they seemed to provide only part of the picture. Kennedy and Miceli (*ibid.*) did not employ any mentalistic measures, so their findings were based solely on the researchers' observation of learners' corpus consultation behaviour, and overlooked the mechanisms underlying such behaviour. Consequently, their generalization of the linear steps of corpus consultation seemed to be more of an idealized 'standard operating procedure' than what actually took place as learners consulted corpora. As evidenced by the examples in Chapter 5, most corpus consultation attempts were by no means as straightforward as Kennedy and Miceli (*ibid.*) prescribed, they were highly discursive, entailing constant changes in strategies (cognitive or metacognitive) and reference resources in the light of instantaneous findings. Whereas Sun (*ibid.*) identified four cognitive skills on which the present findings were based, she did not elaborate on the cognitive skills with learners' accounts of their own endeavours. Again, the linear cognitive skills that Sun (*ibid.*) generalized seemed to be a simplistic reduction of an immensely complex mechanism. In contrast to the two studies discussed above, the present study did not reduce such a cognitively complex endeavour to a simplistic model, but allowed the data to speak for itself, by providing as thick a descriptive account as possible (see instances of corpus consultation endeavours in 5.2 and 5.3). Also, by aligning mentalistic data with behaviouristic data, this study was able to reconstruct what actually took place in the learner's mind as s/he undertook the DALC task.

Broadly speaking, the learners were found to be fairly resourceful in coordinating

various *mediating tools* with their own *mental capacities* to approach the DALC task. A synthesis of SCT and cognitive approaches to SLA accounts for how such endeavours led to *collocation learning* noted in the previous subsection: a corpus consultation endeavour was initiated by the DALC task, which was a culturally constructed auxiliary means of serving pedagogical purposes, followed by the learner employing physical mediating tools such as concordancers to generate real language data (i.e., symbolic mediating tool), which fed into the learner's mind to be processed in working memory by cognitive and metacognitive strategies before it was ready to be committed to long-term memory, hence restructuring collocational knowledge, or *learning* (see cognitive approaches to SLA in 2.4.1 and SCT in 2.4.2). Also, the more intensely the language input was processed, the higher the likelihood it had to be learned. In this sense, the changes that occurred at various levels of learners' collocational knowledge, as illustrated in Chapter 4, can be accounted for by the complex interplay between the physical mediation, symbolic mediation and information processing induced by corpus consultation.

### 7.1.3 Learner perceptions of DALC

RQ3. How do EFL learners perceive the data-driven approach to learning collocations?

Learner perceptions of DALC were elicited through a questionnaire at the closure of the intervention. As regards collocational awareness, *collocation* was perceived to be of only marginal importance as a component of word knowledge, as compared with *word meaning*, *grammatical function*, *spoken and written form* (see 6.3.2). The learners' disregard of collocation may have resulted from the preconception of words as discrete units, overlooking the fact that frequently co-occurring words were also an integral part of word knowledge. Despite the lesser importance relative to other components of word knowledge, collocation was perceived to contribute to the *accuracy*, *fluency* and *appropriateness* of language production, as generally claimed in the literature (see the results in 6.4 and the rationale for developing collocational knowledge in 2.3.1). Recognizing the importance of collocation in language production, the respondents also expressed a willingness to devote more attention to collocations in future English learning. Admittedly, the favourable attitude toward collocation learning may well be the result of *halo effect* (i.e., the phenomenon in



which the respondent attempts to please the researcher by giving responses that are desirable), as is the case with many questionnaire or interview surveys. The discrepancy between the perceived importance of collocation in *language production* and that in *word knowledge* calls for the need to change learners' preconception of vocabulary as discrete units and to develop an awareness of the formulaic nature of language.

The pretest results show that the participants had a rather limited repertoire of collocations (see 4.1.1), although they reported to have paid attention to collocations while learning vocabulary (see 6.3.1). To compound the problem, not only the participants' *knowledge* but also their *reference resources* of collocations were limited: a fairly high proportion of respondents had not used concordancers or collocation dictionaries before the time of the study (see 6.2). It is likely that their disregard of collocation as part of word knowledge prevented them from seeking and/or using reference resources for collocational information. Conversely, it may well be that limited access to reference resources impeded their development of collocational knowledge. Either way, the results indicate that the resources the learners knew for consulting or learning collocations were scarce, suggesting the need to introduce learners to more resources for accessing collocational information.

Having experienced DALC, the participants generally perceived the benefits of corpus consultation. The rationales behind DALC highlighted in Section 2.3.2, namely the *genuineness* and *profusion* of language data and the sense of *learner autonomy*, were recognized by the respondents (see 6.5.1 and 6.5.2). They felt that the major strengths of the pedagogical use of corpora lay in the *genuineness* and *profusion* of language data, as opposed to the artificial, limited examples in textbooks. As far as vocabulary learning was concerned, learners found corpus consultation helpful in *consolidating* and *enhancing* word knowledge (see 6.5.2.1). More specifically, corpus resources were deemed to be most useful for consulting *word use* (e.g., grammatical function and collocation), in contrast to *word meaning* (e.g., meaning and connotation) and *word form* (i.e., written and spoken form). In addition, learners found that the wealth of corpus data not only highlighted the syntagmatic relations of words (i.e., collocation) but also helped in differentiating near-synonyms (i.e., one of the paradigmatic relations of words). Beyond the linguistic level, corpus consultation was

regarded to enhance the sense of *autonomous learning* (see 6.5.2.2).

Reflecting on their hands-on experience of corpus consultation, the participants did not feel corpus data overwhelming in terms of language difficulty and format, but they found corpus size problematic (e.g., too large to generalize collocations or too small to identify a pattern) and the entire consultation process laborious (see 6.5.1). Related to the informants' concern about efficiency was their strong preference for two particular concordancers over the others, namely *Collins WordbanksOnline* and *IWill* (see 6.5.3), because these two tools supported POS search function which efficiently narrowed the search scope. Notably, the participants' reliance on *Collins WordbanksOnline* was evident from the fact that the temporary breakdown of this particular concordancer resulted in the unusually poor performance on one of the DALC tasks (as discussed in 4.3). Overall, the respondents felt confident about the continual use of corpus resources in assisting language learning. This positive perception was reflected in the learners' performance data: a positive correlation ( $r = .197$ ,  $p = .029$ ) was found between the claimed willingness of continual use of concordancers in aiding collocation learning and the extent of progress in collocation tests. In other words, the more progress a learner made in collocation tests, the more willing s/he was to use concordancers for future collocation learning. Alternatively, the association can be interpreted as: the more confident a learner felt about the use of concordancers, the more progress s/he made in collocation tests. The precise causal inference cannot be made, but the statistically significant association nonetheless shed light on the link between learners' performance and perception.

In sum, the questionnaire responses are indicative of a generally positive attitude toward corpus-assisted language and/or collocation learning. Notwithstanding the perceived benefits, learners held reservations about the efficiency of inducing collocations from concordances. Although the time and processing effort invested in such endeavours are arguably what contribute to *learning* (see 2.4.1.2 for the *depth of processing hypothesis*), the concerns about efficiency may nonetheless affect the learner's willingness to consult corpora for language learning purposes. This therefore ties in with Kennedy and Miceli's (2010:41) call for 'preparedness to proceed with trial and error' and O'Sullivan's (2007) proposal for developing 'corpus consultation literacy', in order to guide learners to make the most of corpus resources for language

learning. The pedagogical implications will be discussed in Section 7.3.

## **7.2 Limitations and directions for future research**

The following subsections critically reflect on the limitations of the current research and point to some possible ways forward for future research.

### **7.2.1 Measuring the free production of collocations**

In this study the participants' *free productive knowledge* of collocations was measured by the collocations used in their writing assignments. As noted in Section 3.5.2.2, in order to minimize any disturbances that this study might have caused to the curriculum in which the participants were enrolled, the written production was conveniently sampled from writing assignments, rather than having the learners write a specific piece for the sole purpose of this research. This study was thus unable to impose stringent control on the written production collected: the four sets of writing assignments (i.e., 2 groups x 2 points in time) varied in genre and writing topic (see 3.5.2.2), and the target collocations covered in the intervention sessions were not required to be used in the writing assignments. Unfortunately, such a convenience sampling of written production rendered certain comparisons of collocation use unlikely, if not impossible (e.g., between-group comparisons), because collocation profiles may vary with genre/register, though no empirical evidence has hitherto shown how writing topics determine the range of lexical collocations used therein.

Nonetheless, it was precisely the lesser control on the writing samples that allowed an insight into how DALC extended its impact beyond the knowledge of the target collocations to overall collocational knowledge and awareness, and how such increased knowledge and awareness was reflected on general writings instead of prescribed ones (see the impact of DALC on learners' free production of collocations in 4.2). Whereas the present study found positive changes in the learners' free productive collocational knowledge regardless of different writing genres and topics, it is worthwhile for future research on free collocation production to control the types of writing so as to provide a more solid basis for comparisons of collocation use.

In addition, in this study one of the indicators of the quality of collocation production



is *acceptability* (see 4.2.3). Taking into account the typicality/commonness of a word combination and the context in which it was embedded, *acceptability* here was judged in an acceptable-or-unacceptable, *dichotomous* manner, but as discussed in Section 2.2.3.1, it may be better manifested in a *continuous* manner. Nesselhauf (2003) used a five-point scale to measure the degree of acceptability of a collocation. Schmitt (1998a) used a scale of 0-3 to indicate the learner's ability to use a target word productively. It is therefore suggested that when measuring free productive collocational knowledge in future research, a more comprehensive measurement may be obtained by using multiple indicators (e.g., restrictedness, acceptability) manifested in a continuous manner. To the best of my knowledge, this study is the first attempt to examine the impact of corpus consultation on the *free production* of collocations in addition to *recognition* and *controlled production*, it is hoped that more effort will be devoted to investigating the ways in which corpus consultation influences various levels of collocational knowledge.

## **7.2.2 Additional factors influencing corpus-assisted language learning**

### **1) Task types**

Drawing on my observation of learners' corpus consultation in think-aloud sessions, it seems that three particular types of worksheet question induced more intense cognitive processing: *MC question*, *semantic grid*, and *error identification and correction question* (see Appendix 10 for the think-aloud worksheet). The first two types entailed comparison and contrast among candidate collocates, because the distracters were either semantically similar or literally translated from L1. Such question types were therefore likely to induce more corpus interrogation and deeper information processing to distinguish the target collocations from the distracting ones. Lewis (2000) also suggests contrasting miscollocations with the target ones to raise the awareness of collocability. In addition, the error identification and correction questions also required a higher cognitive involvement on the part of the learner as they activated two levels of collocational knowledge: recognition and controlled production. As tantalizing as they may seem, the assertions above concerning the relationship between question types and the cognitive involvement load induced were

made on the sole basis of my observation of learners' corpus consultation processes. Laufer and Hulstijn (2001) predict in their *involvement load hypothesis* (see 2.4.1.2) that tasks with a higher involvement load will better facilitate vocabulary retention than those with a lower involvement load. This study did not set out to examine the relationships among the types of corpus-driven language task, cognitive involvement load induced and the resulting learning outcomes, but the findings confirmed that a language task as cognitively demanding as DALC did contribute to collocation learning. Such serendipitous observation points to a direction for future investigations: the ways in which and the extent to which collocation learning is mediated by different types of corpus-driven language task.

## **2) Individual differences**

Questionnaire responses illustrate that the experimental participants were comfortable using computer and web-based resources for language learning (see 6.2), and felt confident about their concordancing skills (6.5.1). Notwithstanding the above positive *perceptions* of their own corpus consultation literacy, the *performance* data indicates a considerable variation among learners (as discussed 4.1.2): that is, different learners responded to DALC intervention rather differently despite an overall trend of progress. Although an association was found between learners' vocabulary learning experience and the degree to which they benefited from DALC (see 4.1.2), the present research design was unfortunately not able to further explain the marked variation among learners, but speculated that this may have been the result of individual differences. It therefore merits further inquiries into the role of individual differences in corpus-assisted language learning, in terms of learner orientation toward technology-enhanced language learning, computer literacy, learning style, language proficiency, and so forth.

### **7.2.3 Strengthening the product-process link with case studies**

This study employed a mixed method which combined both *quantitative* and *qualitative* measures. To ensure the external validity (i.e., generalizability) of the *quantitative* datasets (e.g., test, questionnaire), the sample size of the present study was fairly large (N=186). While a large cohort allows the commonality within the sample to emerge, it is difficult, if not impossible, to gain more in-depth insights with

such a large sample. The huge cohort size allowed this study to determine the extent to which corpus consultation mediated collocation learning, but precluded the possibility to probe further into the ways in which each participant fared individually with corpus consultation. Therefore, the *product* data was based on the entire sample (N=186), while the *process* data came from a sub-sample (N=17). Note that owing to the idiosyncratic nature of corpus consultation, the sub-sample was not designed to be representative of the entire sample but only to reveal the potential repository of strategies learners may employ to fare with the DALC task. It is therefore suggested that a stronger link between the *product* and *process* data can be established through small-scale case studies. To understand the product-process links more clearly, future research may address issues such as what specific mental process attributes to the development of which aspect of collocational knowledge. In-depth case studies looking closely into a small group of participants' corpus consultation processes and the corresponding changes in their collocational knowledge may provide more evidence of the causal inferences between corpus consultation endeavours and the learning outcomes.

#### **7.2.4 Triangulating questionnaire with follow-up interview**

A questionnaire was administered to the experimental group after DALC intervention to elicit their perceptions of the corpus-assisted collocation learning experience. Whereas questionnaire responses reflect how these learners perceived such a pedagogical approach in general, more in-depth information thereof could have been elicited by follow-up interviews. For example, questionnaire items were unable to capture instances in which the learner expressed a feeling of frustration, as demonstrated in one of the segments of verbal reports (see 5.3.2.1). A follow-up interview is suggested, to probe further into learners' experience and perceived strengths and limitations of corpus-assisted language learning by asking them to elaborate on their feelings and/or opinions with specific examples.



### 7.2.5 Diachronic investigation into corpus-assisted collocation learning

Due to the practical constraints on the access granted to the participants and the research site, there was only one post-test administered (i.e., an immediate post-test) in the current research. If time permits, future research may incorporate a delayed post-test in addition to an immediate one, in order to see the extent to which the effects of corpus pedagogy persist over time.

Collocation learning is an incremental process: developing collocational knowledge entails not only expanding the *breadth* by establishing more collocational links, but also increasing the *depth* by enhancing the strengths of collocational links, so as to push *receptive knowledge* through *controlled productive knowledge* toward *free productive knowledge*. The present study examined a group of EFL learners' corpus-assisted collocation learning over the course of two months, a relatively short period of time compared with longitudinal studies. Given the short timeframe, there may be limits in demonstrating the full potential of corpus pedagogy on the long-term development of collocational knowledge. Future investigations may focus attention on a specific group of words and observe for an extended period of time to determine the ways in which and the extent to which corpus consultation impacts the *breadth* and *depth* of learners' collocational knowledge in the long term.

## 7.3 Pedagogical implications

This study has 1) borne out the efficacy of DALC on developing collocational knowledge and raising collocational awareness; 2) revealed the extremely complex mechanisms contributing to collocation learning; and 3) explored learner perceptions of corpus-assisted collocation and/or language learning. The findings have important implications for FL pedagogy as to teaching collocations (7.3.1) and scaffolding corpus-assisted language learning (7.3.2).

### 7.3.1 Developing collocational knowledge and raising collocational awareness

As evidenced by the learners' test performance, their *collocational knowledge* lagged far behind *definitional knowledge* of target words (see 4.1.1), which may be attributable to the lesser importance they attached to collocation as part of word knowledge (see the perceived importance of collocation relative to other components of word knowledge in 6.3.2). It may well be that vocabulary has been taught and/or learned as discrete units, so learners have only limited awareness and knowledge of collocations, and even a misconception that relations between words are determined exclusively by semantics (as demonstrated in 5.2.1 by S5's reflection on how she learned the words *urban*, *suburban* and *rural*). It seems that vocabulary pedagogy and learning in this particular EFL context is rather problematic, as evidenced by learners' *performance* and *perceptions* of collocations. This has implications for FL vocabulary pedagogy to raise learners' awareness of the *multidimensionality* of lexical knowledge, and to sensitize them to the less marked aspects such as collocation, word association and connotation (see word knowledge framework in 2.2.1). To put it differently, more pedagogical attention should be devoted to increasing the *depth* of lexical knowledge, of which collocation is a crucial aspect. In addition, it is also imperative to change the learners' and/or teachers' misconception of learning words in isolation, and to redirect their attention beyond the boundary of *word* to *lexical unit*, which conveys a single meaning with one or more than one word (cf. Bogaards, 2001; Cruse, 1986). Learners need to be made aware of *collocational relations* or *restrictions* among words, as they are dictated by actual usage instead of semantics alone. Closely related to the usage-based nature of collocation is its language specificity: collocation is not based exclusively on semantics, nor on cross-linguistic transfer. It is therefore necessary to bring the learners' attention to the language-specific nature of collocation. A rich array of pedagogical activities has been proposed to develop collocational knowledge: comparing and contrasting collocations and miscollocations, comparing and contrasting collocations in L1 and the TL, using a semantic grid, de-lexicalized words matrix, teacher-filtered or learner-access concordancing activities.

### **7.3.2 Scaffolding corpus-assisted language learning**

As the learners' performance data attested to the effectiveness of DALC and their perception data reflected a general positive attitude toward such an approach, it is important to consider the ways by which corpus resources can better serve learning purposes, most notably developing learners' corpus consultation skills, raising critical awareness of corpus resources and corpus consultation, and fostering autonomous learning.

#### **7.3.2.1 Developing corpus consultation skills**

As far as collocation is concerned, reference tools such as dictionaries or thesauri are of limited use, or even misleading (as illustrated in Chapter 5), so web-based corpora can provide a credible alternative for consulting collocational information. For learner-access corpus consultation to be effective, many researchers call for adequate training on corpus consultation skills (Bernardini, 2004; Braun, 2005; Cheng, Warren and Xu, 2003; Kaur and Hegelheimer, 2005; Kennedy and Miceli, 2001, 2010; O'Sullivan, 2007; Stevens, 1991; Sun, 2003; Yoon and Hirvela, 2004), including choosing a corpus which is fit for purpose (e.g., spoken or written, formal or informal), using suitable query functions (e.g., POS search, wildcard search and alphabetical sorting), modifying query words/phrases as necessary, and so on. To scaffold learners in exploiting corpus resources, it is suggested that they may start from smaller, more specific corpora with basic features, and then move onto larger, more comprehensive ones with advanced features as they gain mastery of corpus consultation skills with practice. The aim of developing corpus consultation skills is not so much as to present target collocations, but to equip learners with the skills to independently exploit such resources in their future language learning. It is hoped that learners not only learn *what* is taught by the teacher in the classroom, but also *how* to learn by themselves outside the classroom. After all, acquainting language learners with a broader range of reference resources and the skills necessary to use them, in order to aid their autonomous learning, is one of the principal aims of language education.



### 7.3.2.2 Raising critical awareness of the strengths and limitations of corpus resources

Having explored corpus consultation endeavours in greater depth, I am convinced that in addition to the above skills of physically manipulating language data in corpora, learners also need to have *critical awareness* of the strengths and limitations of corpus resources: what real language data in corpora represents is *probability*, not *certainty*, particularly in the case of collocations. O'Keeffe and Farr (2003) echo the call for a *critical awareness* of what corpus findings represent, and warn against accepting them readily as absolute truths. A corpus is a sample of a language: the larger a corpus is, the more likely it is to represent the tendency of language use. But no matter how large a corpus might be, it is by no means exhaustive in terms of real language use. Therefore, learners should be aware of the fact that even if one corpus does not contain a particular collocation, it does not follow that the collocation does not exist (Owen, 1993), as it may well be the sampling problems or the choice of corpora; on the other hand, even if a collocation occurs frequently in a corpus, it might not necessarily fit into any contexts learners have to hand. A critical view of corpora as reference is important: corpora provide raw language data to be analyzed with caution, not to be adapted blindly. Particularly, in the case of collocation search, consultation of multiple sources is necessary to better inform the *probability* of a target collocation.

The fundamental difference between corpus resources and others (e.g., dictionaries or thesauri) is that the former provides raw materials from which the user induces language patterns, assuming the role of a researcher; the latter presents straightforward linguistic descriptions generalized by professional lexicographers or grammarians. As evidenced by the highly discursive processes discussed in Chapter 5, corpus consultation is clearly characterized by *trial and error*, and is not as straightforward as consulting dictionaries or grammar books. It can be argued that language use is, by its very nature, more varied and less definite than word definitions or syntactic structures, so consulting *language in use* (e.g., real language in corpora) for *language use* (e.g., word use such as collocation) inevitably entails more synthesis and analysis on the part of the learner. In the light of their observation on corpus consultation behaviour, Kennedy and Miceli (2010:41) suggest that 'working with

corpora requires greater preparedness to proceed by trial and error than work with other reference resources, and acceptance of the uncertainty of finding a satisfactory answer'. Corpus consultation is by its nature *inductive*, and thus entails inductive reasoning on the part of the learner, as opposed to many conventional deductive approaches to FL pedagogy. As FL pedagogy in this particular research context is still predominantly deductive, inductive approaches to language learning can be challenging for the learners and therefore necessitate proper guidance on inductive reasoning skills. It is noteworthy that the call above does not go so far as to suggest specific trainings on inductive learning strategies *per se* for the sole purpose of corpus consultation, but to raise the awareness that inductive reasoning merits more pedagogical attention, because it plays an important role in corpus consultation as well as in language learning in general.

### 7.3.2.3 Fostering autonomous learning

Corpus pedagogy is characterized by autonomous learning, as supported by the learner perceptions shown in the questionnaire (see 6.5.2.2). As corpus-assisted learning settings provide learners with mediating tools to access a wealth of the TL in use, the classroom dynamic changes accordingly. Johns (1991) aptly uses the metaphor of the learner as a *researcher* and the teacher as a *director* or *coordinator* of the student-initiated research. Bernardini (2004) goes on to claim that the teacher should act as a *learning expert* instead of a *language expert*, as s/he facilitates learners' linguistic exploration but does not prescribe what to be learned. The findings of this study have implications for the teacher to scaffold learners' corpus-assisted language learning by providing assistance in their quest for knowledge, rather than imparting knowledge directly. Related to the changes in the classroom dynamic is a shift in the *authority* of the TL from the teacher alone to real language in corpora. In a conventional learning setting, the teacher and the textbook are the only authority of the TL, whereas in a corpus-assisted language classroom, the TL can be easily accessed through corpora. This is not to disregard the role of the teacher in the FL classroom, but to empower the teacher, particularly non-native ones, as well as the learner with a wider range and more reliable sources of the TL.

## 7.4 Concluding remarks

With the advent of computer technology, corpus resources have become increasingly prevalent and user-friendly, so it is timely to explore how repositories of language in use can be pedagogically exploited, and how learners can avail themselves of such valuable resources in language learning. The present study investigated an innovative pedagogical approach, DALC, from three dimensions: *learning product*, *learning processes*, and *learner perceptions*. The findings demonstrate that DALC was beneficial for extending EFL learners' *receptive*, *controlled productive* and *free productive knowledge* of lexical collocations. DALC also heightened learners' awareness of the *ubiquitous*, *usage-based* and *language-specific* nature of collocations. The increase in learners' collocational knowledge may be attributable to the intense cognitive processing of corpus data induced by the DALC task, as evidenced by a rich array of cognitive and metacognitive strategies employed by the learners to approach the task. The learners in the study generally held positive attitudes toward DALC in assisting collocation learning, and they found corpus resources particularly useful for consulting information on *word use* (e.g., grammatical function or collocation). However, notwithstanding the perceived benefits, they had reservations about the time and effort needed for accessing collocational information through corpus consultation.

One of the perpetual challenges facing collocation learning in FL contexts is the paucity of naturalistic exposure to the TL necessary for establishing and strengthening collocational links. This therefore necessitates explicit pedagogical actions to develop collocational knowledge and raise collocational awareness, however, formal tuition in many EFL contexts such as this one does not seem to devote due attention in this respect. To compound the problem, even if collocation is attended to in the language classroom, given the ubiquity and variations of collocations in a language, it is impossible to teach all of them in the classroom, as is the case with teaching vocabulary. Hence, FL pedagogy needs to prepare learners with the *skills*, *resources* and *awareness* to learn collocations autonomously. Pedagogical attention should be devoted to raising learners' awareness that vocabulary cannot be learned in isolation: collocation is an integral part of word knowledge, and plays an important role in language comprehension and production. In addition to awareness-raising, learners



need to be guided to exploit available resources judiciously for self-directed collocation learning outside the classroom. Given the prevalence of computers and Internet access (e.g., all respondents reported having access to both at home), web-based corpora and concordancers can provide rich, easily-accessible resources of collocational information. Corpus resources hold considerable promise for facilitating self-directed collocation learning, not only because of the wealth of real language data therein, but also the opportunity they provide for intense cognitive processing of language input (e.g., observation, synthesis and analysis). A critical awareness and adequate skills for consulting repositories of language in use may go a long way toward autonomous collocation learning.

Exploring from multiple dimensions, the present study endeavours to provide a comprehensive picture of how collocation learning is mediated by corpus consultation. Though by no means exhaustive, this study points to a direction in which future research can fruitfully proceed. It is envisaged that this research adds empirical evidence to the growing body of literature on the pedagogical use of corpus resources, as well as providing a solid basis for a wider application of DALC in L2 contexts.

## References

- Aisenstadt, E. (1981) Restricted Collocations in English Lexicology and Lexicography. *I.T.L. Review of Applied Linguistics*, 53, 53-61.
- Aston, G. (1995) Corpora in Language Pedagogy: Matching Theory and Practice. IN COOK, G. & SEIDLHOFER, B. (Eds.) *Principle and Practice in Applied Linguistics: Studies in Honour of H.G. Widdowson*. Oxford: Oxford University Press.
- Baddeley, A. D. (1997) *Human Memory: Theory and Practice*. Hove: Psychology Press.
- Bahns, J. (1993) Lexical Collocations: A Contrastive View. *ELT Journal*, 47(1), 56-63.
- Bahns, J. & Eldaw, M. (1993) Should We Teach EFL Students Collocations? *System*, 21(1), 101-114.
- Bannert, M. & Mengelkamp, C. (2008) Assessment of Metacognitive Skills by Means of Instruction to Think Aloud and Reflect When Prompted. Does the Verbalisation Method Affect Learning? *Metacognition and Learning*, 3(1), 39-58.
- Barlow, M. (1996) Corpora for Theory and Practice. *International Journal of Corpus Linguistics*, 1(1), 1-37.
- Beheydt, L. (1987) The Semantisation of Vocabulary in Foreign Language Learning. *System*, 15(1), 55-67.
- Belz, J. A. & Vyatkina, N. (2008) The Pedagogical Mediation of a Developmental Learner Corpus for Classroom-Based Language Instruction. *Language Learning and Technology*, 12(3), 33-52.
- Benson, M., Benson, E. & Ilson, R. F. (1986) *Lexicographic Description of English*. Philadelphia: John Benjamins.
- Benson, M., Benson, E. & Ilson, R. F. (1997) *The BBI Dictionary of English Word Combinations*. Amsterdam: John Benjamins.
- Bernardini, S. (2000) Systematising Serendipity: Proposals for Concordancing Large Corpora with Language Learners. IN BURNARD, L. & MCENERY, T. (Eds.) *Rethinking Language Pedagogy from a Corpus Perspective*. Frankfurt: Peter Lang.
- Bernardini, S. (2001) Spoilt for Choice: A Learner Explores General Language Corpora. IN ASTON, G. (Ed.) *Learning with Corpora*. Houston, Texas: Athelstan.
- Bernardini, S. (2002) Exploring New Directions for Discovery Learning. *Language and Computers*, 42(1), 165-182.
- Bernardini, S. (2003) Designing a Corpus for Translation and Language Teaching: The CEXI Experience. *TESOL Quarterly*, 37(3), 528-537.
- Bernardini, S. (2004) Corpora in the Classroom - an Overview and Some Reflections on Future Developments. IN SINCLAIR, J. (Ed.) *How to Use Corpora in Language Teaching*. Amsterdam: John Benjamins.
- Berry-Rogghe, G. L. M. (1973) The Computation of Collocations and Their Relevance in Lexical Studies. IN AITKEN, A. J., BAILEY, R. W. & HAMILTON-SMITH, N. (Eds.) *The Computer and Literary Studies*. Edinburgh: Edinburgh University Press.
- Bialystok, E. (1978) A Theoretical Model of Second Language Learning. *Language Learning*, 28(1), 69-83.

- Block, D. (2003) *The Social Turn in Second Language Acquisition*. Edinburgh: Edinburgh University Press.
- Bodker, S. (1997) Computers in Mediated Human Activity. *Mind, culture and activity: An international journal*, 4(3), 149-158.
- Bogaards, P. (2001) Lexical Units and the Learning of Foreign Language Vocabulary. *Studies in Second Language Acquisition*, 23(3), 321-343.
- Bolinger, D. (1976) Meaning and Memory. *Forum Linguisticum*, 1, 1-14.
- Bolinger, D. & Sears, D. (1981) *Aspects of Language*. New York: Harcourt Brace Jovanovich.
- Bonk, W. J. (2001) Testing ESL Learners' Knowledge of Collocations IN HUDSON, T. & BROWN, J. D. (Eds.) *A Focus on Language Test Development: Expanding the Language Proficiency Construct across a Variety of Tests*. Manoa: University of Hawaii.
- Boulton, A. (2009) Testing the Limits of Data-Driven Learning: Language Proficiency and Training. *ReCALL*, 21(1), 37-54.
- Bowker, L. (1999) Exploring the Potential of Corpora for Raising Language Awareness in Student Translators. *Language Awareness*, 8(3-4), 160-173.
- Bowles, M. (2010a) Concurrent Verbal Reports in Second Language Acquisition Research. *Annual Review of Applied Linguistics*, 30(1), 111-127.
- Bowles, M. (2010b) *The Think-Aloud Controversy in Second Language Research*. New York: Routledge.
- Bowles, M. A. & Leow, R. P. (2005) Reactivity and Type of Verbal Report in SLA Research Methodology: Expanding the Scope of Investigation. *Studies in Second Language Acquisition*, 27(3), 415-440.
- Brannen, J. (Ed.) (1992) *Mixing Methods: Qualitative and Quantitative Research*, Aldershot: Ashgate.
- Braun, S. (2005) From Pedagogically Relevant Corpora to Authentic Language Learning Contents. *ReCALL*, 17(1), 47-64.
- Brown, A. L. & Palincsar, A. S. (1982) Inducing Strategic Learning from Texts by Means of Informed, Self-Control Training. *Topics in Learning and Learning Disabilities*, 2(1), 1-17.
- Brown, J. D. (2001) *Using Surveys in Language Programs*. Cambridge: Cambridge University Press.
- Brown, J. D. & Rodgers, T. S. (2002) *Doing Second Language Research*. Oxford: Oxford University Press.
- Bruner, J. S. (1966) On Cognitive Growth. IN BRUNER, J. S., OLVER, R. R. & GREENFIELD, P. M. (Eds.) *Studies in Cognitive Growth*. New York: Wiley.
- Carr, W. (1995) *For Education: Towards Critical Educational Inquiry*. Buckingham, Philadelphia: Open University Press.
- Carter, R. (1988) Vocabulary, Cloze and Discourse: An Applied Linguistic View. IN CARTER, R. & MCCARTHY, M. (Eds.) *Vocabulary and Language Teaching*. London; New York: Longman.
- Carter, R. (1998) *Vocabulary: Applied Linguistic Perspectives*. New York: Routledge.
- Carter, R. & McCarthy, M. (1995) Grammar and the Spoken Language. *Applied Linguistics*, 16(2), 141-158.
- Chambers, A. (2005) Integrating Corpus Consultation in Language Studies. *Language Learning & Technology*, 9(2), 111-125.
- Chambers, A. & O'Sullivan, I. (2004) Corpus Consultation and Advanced Learners' Writing Skills in French. *ReCALL*, 16(1), 158-172.
- Chamot, A. U. (2004) Issues in Language Learning Strategy Research and Teaching.



- Electronic Journal of Foreign Language Teaching*, 1(1), 14-26.
- Chamot, A. U., Kupper, L. & Impink-Hernandez, M. V. (1988) *A Study of Learning Strategies in Foreign Language Instruction: Finding of the Longitudinal Study*. McLean, Va.: Interstate Research Associates.
- Chan, T. P. & Liou, H. C. (2005) Effects of Web-Based Concordancing Instruction on EFL Students' Learning of Verb-Noun Collocations. *Computer Assisted Language Learning*, 18(3), 231-250.
- Chang, W. C. (1997) *Freshmen English Composition: An Error Analysis from the Discourse Perspective*. Taipei: Crane.
- Chapelle, C. A. (2001) *Computer Applications in Second Language Acquisition: Foundations for Teaching, Testing and Research*. Cambridge: Cambridge University Press.
- Charles, M. (2007) Reconciling Top-Down and Bottom-up Approaches to Graduate Writing: Using a Corpus to Teach Rhetorical Functions. *Journal of English for Academic Purposes*, 6(4), 289-302.
- Cheng, W., Warren, M. & Xu, X. F. (2003) The Language Learner as Language Researcher: Putting Corpus Linguistics on the Timetable. *System*, 31(2), 173-186.
- Chenoweth, N. A. (1995) Formulaicity in Essay Exam Answers. *Language Sciences*, 17(3), 283-297.
- Chomsky, N. (1995) *The Minimalist Program*. Cambridge, MA: MIT Press.
- Cobb, T. (1997) Is There Any Measurable Learning from Hands-on Concordancing? *System*, 25(3), 301-315.
- Cobb, T. (1999) Breadth and Depth of Lexical Acquisition with Hands-on Concordancing. *Computer Assisted Language Learning*, 12(4), 345-360.
- Cohen, A. D. (1996) Verbal Reports as a Source of Insights into Second Language Learner Strategies. *Applied Language Learning*, 7(1 & 2), 5-24.
- Cohen, A. D. (1998) *Strategies in Learning and Using a Second Language*. Harlow: Longman.
- Cohen, A. D. & Macaro, E. (Eds.) (2007) *Language Learner Strategies: Thirty Years of Research and Practice*. Oxford: Oxford University Press.
- Cohen, L., Manion, L. & Morrison, K. (2007) *Research Methods in Education*. London: Routledge.
- Confrey, J. (1990) What Constructivism Implies for Teaching. *Journal for Research in Mathematics Education. Monograph*, 4, 107-210.
- Conklin, K. & Schmitt, N. (2008) Formulaic Sequences: Are They Processed More Quickly Than Nonformulaic Language by Native and Nonnative Speakers? *Applied Linguistics*, 29(1), 72-89.
- Coulmas, F. (1979) On the Sociolinguistic Relevance of Routine Formulae. *Journal of Pragmatics*, 3(3-4), 239-266.
- Cowie, A. P. (1981) The Treatment of Collocations and Idioms in Learners' Dictionaries. *Applied Linguistics*, 2(3), 223-235.
- Cowie, A. P. (1992) Multiword Lexical Units and Communicative Language Teaching. IN ARNAUD, P. & BEJOINT, H. (Eds.) *Vocabulary and Applied Linguistics*. London: Macmillan.
- Coxhead, A. (2000) A New Academic Word List. *TESOL Quarterly*, 34(2), 213-238.
- Craik, F. I. M. & Lockhart, R. S. (1972) Levels of Processing: A Framework for Memory Research. *Journal of Verbal Learning and Verbal Behavior*, 11(6), 671-684.
- Craik, F. I. M. & Tulving, E. (1975) Depth of Processing and the Retention of Words

- in Episodic Memory. *Journal of Experimental Psychology*, 104(3), 268-294.
- Crotty, M. (1998) *Foundations of Social Research: Meaning and Perspective in the Research Process*. London: Sage.
- Cruse, D. (1986) *Lexical Semantics*. Cambridge: Cambridge University Press.
- Davies, A. (2003) *The Native Speaker: Myth and Reality*. Clevedon: Multilingual Matters.
- De Bot, K., Paribakht, T. S. & Wesche, M. B. (1997) Toward a Lexical Processing Model for the Study of Second Language Vocabulary Acquisition. *Studies in Second Language Acquisition*, 19(3), 309-329.
- De Cock, S. (1998) A Recurrent Word Combination Approach to the Study of Formulae in the Speech of Native and Non-Native Speakers of English. *International Journal of Corpus Linguistics*, 3(1), 59-80.
- Dörnyei, Z. (1995) On the Teachability of Communication Strategies. *TESOL Quarterly*, 29(1), 55-85.
- Dörnyei, Z. (2003) *Questionnaires in Second Language Research: Construction, Administration, and Processing*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Dörnyei, Z. (2007) *Research Methods in Applied Linguistics*. Oxford: Oxford University Press.
- Duffy, T. M. & Cunningham, D. J. (1996) Constructivism: Implications for the Design and Delivery of Instruction. IN JONASSEN, D. H. (Ed.) *Handbook of Research for Educational Communications and Technology*. New York: Simon & Schuster Macmillan.
- Ellis, N. C. (1995) The Psychology of Foreign Language Vocabulary Acquisition: Implications for Call. *Computer Assisted Language Learning*, 8(2), 103-128.
- Ellis, N. C. (1996) Sequencing in SLA: Phonological Memory, Chunking, and Points of Order. *Studies in Second Language Acquisition*, 18(1), 91-126.
- Ellis, N. C. (1997) Vocabulary Acquisition: Word Structure, Collocation, Word-Class, and Meaning. IN SCHMITT, N. & MCCARTHY, M. (Eds.) *Vocabulary: Description, Acquisition and Pedagogy*. Cambridge: Cambridge University Press.
- Ellis, N. C. (1998) Emergentism, Connectionism and Language Learning. *Language Learning*, 48(4), 631-664.
- Ellis, N. C. (2002) Frequency Effects in Language Processing. *Studies in Second Language Acquisition*, 24(2), 143-188.
- Ellis, N. C. (2003) Constructions, Chunking, and Connectionism: The Emergence of Second Language Structure. IN DOUGHTY, C. & LONG, M. H. (Eds.) *The Handbook of Second Language Acquisition*. Malden, MA: Blackwell.
- Ellis, N. C. & Laporte, N. (1997) Contexts of Acquisition: Effects of Formal Instruction and Naturalistic Exposure on Second Language Acquisition. IN de GROOT, A. M. B. & KROLL, J. F. (Eds.) *Tutorials in Bilingualism: Psycholinguistic Perspectives*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Ellis, N. C. & Schmidt, R. (1997) Morphology and Longer Distance Dependencies: Laboratory Research Illuminating the A in SLA. *Studies in Second Language Acquisition*, 19, 145-171.
- Ellis, R. (2003) *Task-Based Language Learning and Teaching*. Oxford: Oxford University Press.
- Ericsson, K. A. & Simon, H. A. (1987) Verbal Reports on Thinking. IN FAERCH, C. & KASPER, G. (Eds.) *Introspection in Second Language Research*. Clevedon; Philadelphia: Multilingual Matters.
- Ericsson, K. A. & Simon, H. A. (1993) *Protocol Analysis: Verbal Reports as Data*.



- Cambridge, Mass.; London: MIT Press.
- Erman, B. & Warren, B. (2000) The Idiom Principle and the Open-Choice Principle. *Text*, 20(1), 29-62.
- Færch, C. & Kasper, G. (1987) From Product to Process: Introspective Methods in Second Language Research. IN FAERCH, C. & KASPER, G. (Eds.) *Introspection in Second Language Research*. Clevedon; Philadelphia: Multilingual Matters.
- Færch, C., Haastруп, K. & Phillipson, R. (1984) *Learner Language and Language Learning*. Clevedon: Multilingual Matters.
- Farghal, M. & Obiedat, H. (1995) Collocations: A Neglected Variable in EFL. *IRAL*, 33(4), 315-331.
- Farr, F. (2008) Evaluating the Use of Corpus-Based Instruction in a Language Teacher Education Context: Perspectives from the Users. *Language Awareness*, 17(1), 25-43.
- Fernando, C. & Flavell, R. (1981) *On Idiom: Critical Views and Perspectives*. Exeter: University of Exeter.
- Field, A. (2005) *Discovering Statistics Using SPSS: (and Sex, Drugs and Rock 'N' Roll)*. London: Sage.
- Firth, J. R. (1957) Modes of Meaning. IN FIRTH, J. R. (Ed.) *Papers in Linguistics, 1934-1951*. London: Oxford University Press.
- Firth, A. & Wagner, J. (2007) Second/Foreign Language Learning as a Social Accomplishment: Elaborations on a Reconceptualized SLA. *Modern Language Journal*, 91, 800-819.
- Fitz-Gibbon, C. T. (1996) *Monitoring Education: Indicators, Quality and Effectiveness*. London: Cassell.
- Flavell, J. H. (1979) Metacognition and Cognitive Monitoring: A New Area of Cognitive Developmental Inquiry. *American Psychologist*, 34, 906-911.
- Flavell, J. H. (1981) Cognitive Monitoring. IN DICKSON, W. P. (Ed.) *Children's Oral Communication Skills*. New York: Academic Press.
- Fleiss, J. L. (1981) *Statistical Methods for Rates and Proportions*. New York: John Wiley.
- Flowerdew, J. (1996) Concordancing in Language Learning. IN PENNINGTON, M. C. (Ed.) *The Power of CALL*. Houston, Texas: Athelstan.
- Flowerdew, J. (2001) Concordancing as a Tool in Course Design. IN GHADESSY, M., HENRY, A. & ROSEBERRY, R. (Eds.) *Small Corpus Studies and ELT: Theory and Practice*. Amsterdam: John Benjamins.
- Ford, N. & Chen, S. Y. (2000) Individual Differences, Hypermedia Navigation, and Learning: An Empirical Study. *Journal of Educational Multimedia and Hypermedia*, 9(4), 281-311.
- Foster, P. (2001) Rules and Routines: A Consideration of Their Role in the Task-Based Language Production of Native and Non-Native Speakers. IN BYGATE, M., SKEHAN, P. & SWAIN, M. (Eds.) *Researching Pedagogic Tasks: Second Language Learning, Teaching and Testing*. London, New York: Longman.
- Fox, G. (1998) Using Corpus Data in the Classroom. IN TOMLINSON, B. (Ed.) *Materials Development in Language Teaching*. Cambridge: Cambridge University Press.
- Frankfort-Nachmias, C. & Nachmias, D. (1992) *Research Methods in the Social Sciences*. London: Edward Arnold.
- Fuentes, A. C. (2001) Lexical Behaviour in Academic and Technical Corpora: Implications for ESP Development. *Language Learning & Technology*, 5(3),



106-129.

- Gable, S. (2001) Over-Indulgence and Under-Representation in Interlanguage: Reflections on the Utilization of Concordancers in Self-Directed Foreign Language Learning. *Computer Assisted Language Learning*, 14(3-4), 269-288.
- Gabrys-Biskup, D. (1992) L1 Influence on Learners' Renderings of English Collocations: A Polish/German Empirical Study. IN ARNAUD, P. & BEJOINT, H. (Eds.) *Vocabulary and Applied Linguistics*. London: Macmillan.
- Gaskell, D. & Cobb, T. (2004) Can Learners Use Concordance Feedback for Writing Errors? *System*, 32(3), 301-319.
- Gass, S. & Mackey, A. (2000) *Stimulated Recall Methodology in Second Language Research*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Gass, S. & Selinker, L. (2008) *Second Language Acquisition: An Introductory Course*. New York; London: Routledge.
- Gavioli, L. & Aston, G. (2001) Enriching Reality: Language Corpora in Language Pedagogy. *ELT Journal*, 55(3), 238-246.
- Ghadessy, M. (1989) The Use of Vocabulary and Collocations in the Writing of Primary School Students in Singapore. *AILA Review*, 6, 110-117.
- Gilmore, A. (2007) Authentic Materials and Authenticity in Foreign Language Learning. *Language Teaching*, 40(2), 97-118.
- Glaser, R. (1988) The Grading of Idiomaticity as a Presupposition for a Taxonomy of Idioms. IN HULLEN, W. & SCHULZE, R. (Eds.) *Understanding the Lexicon Meaning, Sense and World Knowledge in Lexical Semantics*. Tübingen: Max Niemeyer.
- Glaserfeld, E. V. (1987) Learning as a Constructive Activity. IN JANVIER, C. (Ed.) *Problems of Representation in the Teaching and Learning of Mathematics*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Glaserfeld, E. V. (1989) Cognition, Construction of Knowledge, and Teaching. *Synthese*, 80(1), 121-140.
- Granger, S. (1998) Prefabricated Patterns in Advanced EFL Writing: Collocations and Formulae. IN COWIE, A. P. (Ed.) *Phraseology: Theory, Analysis and Applications*. Oxford: Oxford University Press.
- Granger, S. (2002) A Bird's-Eye View of Learner Corpus Research. IN GRANGER, S., HUNG, J. & PETCH-TYSON, S. (Eds.) *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam: John Benjamins.
- Grenfell, M. & Harris, V. (1999) *Modern Languages and Learning Strategies: In Theory and Practice*. London: Routledge.
- Guba, E. G. & Lincoln, Y. S. (1994) Competing Paradigms in Qualitative Research. IN DENZIN, N. & LINCOLN, Y. S. (Eds.) *Handbook of Qualitative Research*. Thousand Oaks, CA: Sage.
- Haastруп, K. (1991) *Lexical Inferencing Procedures or Talking About Words: Receptive Procedures in Foreign Language Learning with Special Reference to English*. Tübingen, Germany: Gunter Narr.
- Halliday, M. A. K. (1985) *Spoken and Written Language*. Oxford: Oxford University Press.
- Harley, B. (1995) The Lexicon in Language Research. IN HARLEY, B. (Ed.) *Lexical Issues in Language Learning*. Amsterdam: John Benjamins.
- Henriksen, B. (1999) Three Dimensions of Vocabulary Development. *Studies in Second Language Acquisition*, 21(2), 303-317.
- Hill, J. (2000) Revising Priorities: From Grammatical Failure to Collocational

- Success. IN LEWIS, M. (Ed.) *Teaching Collocation: Further Developments in the Lexical Approach*. Hove: Language Teaching Publications.
- Hindmarsh, R. (1980) *Cambridge English Lexicon*. Cambridge: Cambridge University Press.
- Howarth, P. (1998a) Phraseology and Second Language Proficiency. *Applied Linguistics*, 19(1), 24-44.
- Howarth, P. (1998b) The Phraseology of Learners' Academic Writing. IN COWIE, A. P. (Ed.) *Phraseology: Theory, Analysis and Applications*. Oxford: Oxford University Press.
- Hunston, S. (2002) *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.
- Jacobs, G. M. & Farrell, T. M. (2001) Paradigm Shift: Understanding and Implementing Change in Second Language Education *TESL-EJ*.
- Johns, T. (1986) Micro-Concord: A Language-Learners Research Tool. *System*, 14(2), 151-162.
- Johns, T. (1991) Should You Be Persuaded: Two Samples of Data-Driven Learning Materials. IN JOHNS, T. & KING, P. (Eds.) *Classroom Concordancing*. Birmingham: Centre for English Language Studies, University of Birmingham.
- Johns, T. (1994) From Printout to Handout: Grammar and Vocabulary Teaching in the Context of Data-Driven Learning. IN ODLIN, T. (Ed.) *Perspectives on Pedagogical Grammar*. Cambridge: Cambridge University Press.
- Jonassen, D. H. (1991) Evaluating Constructivist Learning. *Educational Technology*, 31(9), 28-33.
- Jourdenais, R. (2001) Cognition, Instruction and Protocol Analysis. IN ROBINSON, P. (Ed.) *Cognition and Second Language Instruction*. Cambridge: Cambridge University Press.
- Kaltenbock, G. & Mehlmauer-Larcher, B. (2005) Computer Corpora and the Language Classroom: On the Potential and Limitations of Computer Corpora in Language Teaching. *ReCALL*, 17(1), 65-84.
- Kaur, J. & Hegelheimer, V. (2005) ESL Students' Use of Concordance in the Transfer of Academic Word Knowledge: An Exploratory Study. *Computer Assisted Language Learning*, 18(4), 287-310.
- Kennedy, C. & Miceli, T. (2001) An Evaluation of Intermediate Students' Approaches to Corpus Investigation. *Language Learning & Technology*, 5(3), 77-90.
- Kennedy, C. & Miceli, T. (2010) Corpus-Assisted Creative Writing: Introducing Intermediate Italian Learners to a Corpus as a Reference Resource. *Language Learning & Technology*, 14(1), 28-44.
- Kennedy, G. (2003) Amplifier Collocations in the British National Corpus: Implications for English Language Teaching. *TESOL Quarterly*, 37(3), 467-487.
- Kirk, J. & Miller, M. L. (1986) *Reliability and Validity in Qualitative Research*. Newbury Park: Sage.
- Lantolf, J. P. (2000) Introducing Sociocultural Theory. IN LANTOLF, J. P. (Ed.) *Sociocultural Theory and Second Language Learning*. Oxford: Oxford University Press.
- Lantolf, J. P. & Appel, G. (1994) Theoretical Framework: An Introduction to Vygotskian Approaches to Second Language Research. IN LANTOLF, J. P. & APPEL, G. (Eds.) *Vygotskian Approaches to Second Language Research*. Norwood, NJ: Ablex.



- Lantolf, J. P. & Thorne, S. L. (2006) *Sociocultural Theory and the Genesis of Second Language Development*. Oxford: Oxford University Press.
- Lantolf, J. P. & Thorne, S. L. (2007) Sociocultural Theory and Second Language Learning. IN VANPATTEN, B. & WILLIAMS, J. (Eds.) *Theories in Second Language Acquisition: An Introduction*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Laufer, B. (1997) What's in a Word That Makes It Hard or Easy: Some Intralexical Factors That Affect the Learning of Words. IN SCHMITT, N. & MCCARTHY, M. (Eds.) *Vocabulary: Description, Acquisition and Pedagogy*. Cambridge: Cambridge University Press.
- Laufer, B. (1998) The Development of Passive and Active Vocabulary in a Second Language: Same or Different? *Applied Linguistics*, 19(2), 255-271.
- Laufer, B. & Hulstijn, J. (2001) Incidental Vocabulary Acquisition in a Second Language: The Construct of Task-Induced Involvement. *Applied Linguistics*, 22(1), 1-26.
- Lee, D. & Swales, J. (2006) A Corpus-Based EAP Course for NNS Doctoral Students: Moving from Available Specialized Corpora to Self-Compiled Corpora. *English for Specific Purposes*, 25(1), 56-75.
- Leech, G. & Candlin, C. (1986) *Computers in English Language Teaching and Research*. London: Longman.
- Leow, R. P. & Morgan-Short, K. (2004) To Think Aloud or Not to Think Aloud - the Issue of Reactivity in SLA Research Methodology. *Studies in Second Language Acquisition*, 26(1), 35-57.
- Lewis, M. (2000) *Teaching Collocation: Further Developments in the Lexical Approach*. Hove: Language Teaching Publications.
- Lightbown, P. M. & Spada, N. (2006) *How Languages Are Learned*. Oxford: Oxford University Press.
- Lincoln, Y. S. (1990) The Making of a Constructivist: A Remembrance of Transformations Past. IN GUBA, E. G. (Ed.) *The Paradigm Dialog*. Newbury Park, CA: Sage.
- Macaro, E. (2006) Strategies for Language Learning and for Language Use: Revising the Theoretical Framework. *Modern Language Journal*, 90(3), 320-337.
- Mackey, A. & Gass, S. (2005) *Second Language Research: Methodology and Design*. Mahwah, NJ: Lawrence Erlbaum Associates.
- MacWhinney, B. (1992) Transfer and Competition in Second Language Learning. *Advances in Psychology*, 83, 371-390.
- Martin, M. (1984) Advanced Vocabulary Teaching: The Problem of Synonyms. *Modern Language Journal*, 68(2), 130-137.
- Matsumoto, K. (1994) Introspection, Verbal Reports and Second Language Learning Strategy Research. *Canadian Modern Language Review*, 5(2), 363-386.
- McCarthy, M. (1990) *Vocabulary*. Oxford: Oxford University Press.
- McCloskey, M. L. & Stack, L. (2004) *Visions B: Language, Literature, Content*. Boston: Thomson Heinle.
- McEnery, T. & Wilson, A. (1997) Teaching and Language Corpora (TALC). *ReCALL*, 9(1), 5-14.
- McIntosh, A. (1961) Patterns and Ranges. *Language*, 37(3), 325-337.
- Meara, P. (2009) *Connected Words: Word Associations and Second Language Vocabulary Acquisition*. Amsterdam: John Benjamins.
- Medina, A. (2008) *Concurrent Verbalization, Task Complexity, and Working Memory: Effects on L2 Learning in a Computerized Task*. Unpublished doctoral



- dissertation. Georgetown University.
- Mel'cuk, I. (1998) Collocations and Lexical Functions. IN COWIE, A. P. (Ed.) *Phraseology: Theory, Analysis, and Applications*. Oxford: Oxford University Press.
- Meunier, F. (2002) The Pedagogical Value of Native and Learner Corpora in EFL Grammar Teaching. IN GRANGER, S., HUNG, J. & PETCH-TYSON, S. (Eds.) *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam: John Benjamins.
- Mitchell, R. & Myles, F. (2004) *Second Language Learning Theories*. London: Arnold.
- Moon, R. (1997) Vocabulary Connections: Multi-Word Items in English. IN SCHMITT, N. & MCCARTHY, M. (Eds.) *Vocabulary: Description, Acquisition and Pedagogy*. Cambridge: Cambridge University Press.
- Moon, R. (1998) *Fixed Expressions and Idioms in English: A Corpus-Based Approach*. Oxford: Oxford University Press.
- Mouly, G. J. (1978) *Educational Research: The Art and Science of Investigation*. Boston: Allyn and Bacon.
- Naiman, N., Fröhlich, M., Stern, H. H. & Todesco, A. (1978) *The Good Language Learner*. Toronto: Ontario Institute for Studies in Education, The Modern Language Centre.
- Nassaji, H. (2003) L2 Vocabulary Learning from Context: Strategies, Knowledge Sources, and Their Relationship with Success in L2 Lexical Inferencing. *TESOL Quarterly*, 37(4), 645-670.
- Nation, I. S. P. (1990) *Teaching and Learning Vocabulary*. Massachusetts: Newbury House.
- Nation, I. S. P. (2001) *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press.
- Nattinger, J. R. & Decarrico, J. D. (1992) *Lexical Phrases and Language Teaching*. Oxford: Oxford University Press.
- Neisser, U. (1992) Two Themes in the Study of Cognition. IN PICK, H. L., BROEK, P. V. D. & KNILL, D. C. (Eds.) *Cognition: Conceptual and Methodological Issues*. Washington, DC: American Psychological Association.
- Nesselhauf, N. (2003) The Use of Collocations by Advanced Learners of English and Some Implications for Teaching. *Applied Linguistics*, 24(2), 223-242.
- Nesselhauf, N. (2004) *Collocations in a Learner Corpus*. Amsterdam: John Benjamins.
- Nunan, D. (1992) *Research Methods in Language Learning*. Cambridge: Cambridge University Press.
- O'Keeffe, A. & Farr, F. (2003) Using Language Corpora in Initial Teacher Education: Pedagogic Issues and Practical Applications. *TESOL Quarterly*, 37(3), 389-418.
- O'Malley, J. M. & Chamot, A. U. (1990) *Learning Strategies in Second Language Acquisition*. Cambridge: Cambridge University Press.
- O'Malley, J. M., Chamot, A. U., Stewner-Manzanares, G., Russo, R. P. & Kupper, L. (1985) Learning Strategy Applications with Students of English as a Second Language. *TESOL Quarterly*, 19(3), 557-584.
- O'Sullivan, I. (2007) Enhancing a Process-Oriented Approach to Literacy and Language Learning: The Role of Corpus Consultation Literacy. *ReCALL*, 19(03), 269-286.
- O'Sullivan, I. & Chambers, A. (2006) Learners' Writing Skills in French: Corpus

- Consultation and Learner Evaluation. *Journal of Second Language Writing*, 15(1), 49-68.
- Owen, C. (1993) Corpus-Based Grammar and the Heineken Effect: Lexico-Grammatical Description for Language Learners. *Applied Linguistics*, 14(2), 167-187.
- Oxford, R. (1990) *Language Learning Strategies: What Every Teacher Should Know*. New York: Newbury House.
- Oxford, R. (1999) Learning Strategies. IN SPOLSKY, B. (Ed.) *Concise Encyclopedia of Educational Linguistics*. Oxford: Elsevier.
- Oxford, R., Cho, Y., Leung, S. & Kim, H-J. (2004) Effect of the Presence and Difficulty of Task on Strategy Use: An Exploratory Study. *International Review of Applied Linguistics*, 42(1), 1-47.
- Palmer, H. E. (1933) *Second Interim Report on English Collocations*. Tokyo: Kaitakusha.
- Pawley, A. & Syder, F. H. (1983) Two Puzzles for Linguistic Theory: Nativelike Selection and Nativelike Fluency. IN RICHARDS, J. C. & SCHMIDT, R. W. (Eds.) *Language and Communication*. New York: Longman.
- Peters, A. (1983) *The Units of Language Acquisition*. Cambridge: Cambridge University Press.
- Porte, G. K. (2002) *Appraising Research in Second Language Learning : A Practical Approach to Critical Analysis of Quantitative Research*. Amsterdam: John Benjamins.
- Qian, D. D. (2005) Demystifying Lexical Inferencing: The Role of Aspects of Vocabulary Knowledge. *TESL Canada Journal*, 22(2), 34-54.
- Qiao, H. L. & Sussex, R. (1996) Using the Longman Mini-Concordancer on Tagged and Parsed Corpora, with Special Reference to their Use as an Aid to Grammar Learning. *System*, 24(1), 41-64.
- Randall, M. (2007) *Memory, Psychology and Second Language Learning*. Amsterdam: John Benjamins.
- Read, J. & Nation, P. (2004) Measurement of Formulaic Sequences. IN SCHMITT, N. (Ed.) *Formulaic Sequences: Acquisition, Processing and Use*. Amsterdam: John Benjamins.
- Richards, J. C. (1976) The Role of Vocabulary Teaching. *TESOL Quarterly*, 10(1), 77-90.
- Rubin, J. (1975) What the 'Good Language Learner' Can Teach Us. *TESOL Quarterly*, 9(1), 41-51.
- Rubin, J. (1981) Study of Cognitive Processes in Second Language Learning. *Applied Linguistics*, 11(2), 117-131.
- Ruschoff, B. & Ritter, M. (2001) Technology-Enhanced Language Learning: Construction of Knowledge and Template-Based Learning in the Foreign Language Classroom. *Computer Assisted Language Learning*, 14(3), 219-232.
- Schmidt, R. W. (1990) The Role of Consciousness in Second Language Learning. *Applied Linguistics*, 11(2), 129-158.
- Schmitt, N. (1998a) Measuring Collocational Knowledge: Key Issues and an Experimental Assessment Procedure. *I.T.L. Review of Applied Linguistics*, 119-120, 27-47.
- Schmitt, N. (1998b) Tracking the Incremental Acquisition of Second Language Vocabulary: A Longitudinal Study. *Language Learning*, 48(2), 281-317.
- Schmitt, N. (2000) *Vocabulary in Language Teaching*. Cambridge: Cambridge University Press.



- Schmitt, N. (2008) Instructed Second Language Vocabulary Learning. *Language Teaching Research*, 12(3), 329-363.
- Schmitt, N. & Carter, R. (2004) Formulaic Sequences in Action. IN SCHMITT, N. (Ed.) *Formulaic Sequences: Acquisition, Processing and Use*. Amsterdam: John Benjamins.
- Schmitt, N. & Meara, P. (1997) Researching Vocabulary through a Word Knowledge Framework. *Studies in Second Language Acquisition*, 19(1), 17-36.
- Schrag, F. (1992) In Defense of Positivist Research Paradigms. *Educational Researcher*, 21(5), 5-8.
- Sealey, A. & Thompson, P. (2007) Corpus, Concordance, Classification: Young Learners in the L1 Classroom. *Language Awareness*, 16(3), 208-223.
- Seidlhofer, B. (2005) English as a Lingua Franca. *ELT Journal*, 59(4), 339-341.
- Sinclair, J. (1991) *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
- Sinclair, J. (Ed.) (2004) *How to Use Corpora in Language Teaching*, Amsterdam: John Benjamins.
- Siyanova, A. & Schmitt, N. (2008) L2 Learner Production and Processing of Collocation: A Multi-Study Perspective. *Canadian Modern Language Review*, 64(3), 429-458.
- Skehan, P. (1998) *A Cognitive Approach to Language Learning*. Oxford: Oxford University Press.
- Smadja, F. A. (1989) Lexical Co-Occurrence: The Missing Link. *Literary and Linguistic Computing*, 4(3), 163-168.
- Smith, T. W. (1984) Nonattitudes: A Review and Evaluation. IN TURNER, C. F. & MARTIN, E. (Eds.) *Surveying Subjective Phenomena*. New York: Russell Sage.
- Sokolik, M. E. & Smith, M. E. (1992) Assignment of Gender to French Nouns in Primary and Secondary Language: A Connectionist Model. *Second Language Research*, 8(1), 39-58.
- St. John, E. (2001) A Case for Using a Parallel Corpus and Concordancer for Beginners of a Foreign Language. *Language Learning & Technology*, 5(3), 185-203.
- Stern, H. H. (1975) What Can We Learn from the Good Language Learner? *Canadian Modern Language Review*, 31(4), 304-318.
- Stern, H. H. (1992) *Issues and Options in Language Teaching*. Oxford: Oxford University Press.
- Stevens, V. (1991) Classroom Concordancing: Vocabulary Materials Derived from Relevant, Authentic Text. *English for Specific Purposes*, 10(1), 35-46.
- Stratman, J. F. & Hamp-Lyons, L. (1994) Reactivity in Concurrent Think-Aloud Protocols. IN SMAGORINSKY, P. (Ed.) *Speaking About Writing: Reflections on Research Methodology*. Thousand Oaks, CA: Sage.
- Stubbs, M. (1995a) Collocations and Semantic Profiles: On the Cause of the Trouble with Quantitative Studies. *Functions of Language*, 2(1), 1-33.
- Stubbs, M. (1995b) Corpus Evidence for Norms of Lexical Collocation. IN COOK, G. & SEIDLHOFER, B. (Eds.) *Principle and Practice in Applied Linguistics: Studies in Honour of H.G. Widdowson*. Oxford: Oxford University Press.
- Sun, Y. C. (2003) Learning Process, Strategies and Web-Based Concordancers: A Case Study. *British Journal of Educational Technology*, 34(5), 601-613.
- Sun, Y. C. (2007) Learner Perceptions of a Concordancing Tool for Academic Writing. *Computer Assisted Language Learning*, 20(4), 323-343.



- Sun, Y. C. & Wang, L. Y. (2003) Concordancers in the EFL Classroom: Cognitive Approaches and Collocation Difficulty. *Computer Assisted Language Learning*, 16(1), 83-94.
- Thompson, P. & Tribble, C. (2001) Looking at Citations: Using Corpora in English for Academic Purposes. *Language Learning & Technology*, 5(3), 91-105.
- Thurstun, J. & Candlin, C. N. (1998) Concordancing and the Teaching of the Vocabulary of Academic English. *English for Specific Purposes*, 17(3), 267-280.
- Todd, R. W. (2001) Induction from Self-Selected Concordances and Self-Correction. *System*, 29(1), 91-102.
- Tribble, C. (1997) Improvising Corpora for ELT: Quick-and-Dirty Ways of Developing Corpora for Language Teaching. IN LEWANDOWSKA-TOMASZCZYK, B. & MELIA, P. J. (Eds.) *Practical Applications in Language Corpora PALC 97*. Lodz: Lodz University Press.
- Tribble, C. & Jones, G. (1990) *Concordances in the Classroom*. Harlow: Longman.
- Tsui, A. B. M. (2004) What Teachers Have Always Wanted to Know - and How Corpora Can Help. IN SINCLAIR, J. (Ed.) *How to Use Corpora in Language Teaching*. Amsterdam: John Benjamins.
- Underwood, G., Schmitt, N. & Galpin, A. (2004) The Eyes Have It: An Eye-Movement Study into the Processing of Formulaic Sequences. IN SCHMITT, N. (Ed.) *Formulaic Sequences: Acquisition, Processing and Use*. Amsterdam: John Benjamins.
- Van Lier, L. (2000) From Input to Affordance: Social-Interactive Learning from an Ecological Perspective. IN LANTOLF, J. P. (Ed.) *Sociocultural Theory and Second Language Learning*. Oxford: Oxford University Press.
- Van Roey, J. (1990) *French-English Contrastive Lexicology: An Introduction*. Louvain-la-Neuve: Peeters.
- Van Someren, M. W., Barnard, Y. F. & Sandberg, J. A. C. (1994) *The Think Aloud Method: A Practical Guide to Modelling Cognitive Processes*. London: Academic Press.
- Vygotsky, L. S. (1978) *Mind in Society: The Development of Higher Psychological Processes*. Cambridge, MA: Harvard University Press.
- Vygotsky, L. S. (1981) The Instrumental Method in Psychology. IN WERTSCH, J. V. (Ed.) *The Concept of Activity in Soviet Psychology*. Armonk, NY: ME Sharpe.
- Vygotsky, L. S. (1997) *The Collected Works of L. S. Vygotsky. Volume 4. The History of the Development of Higher Mental Functions*. London, New York: Plenum.
- Wade, S. E. (1990) Using Think Alouds to Assess Comprehension. *The Reading Teacher*, 43(7), 442-451.
- Weinert, R. (1995) The Role of Formulaic Language in Second Language Acquisition: A Review. *Applied Linguistics*, 16(2), 180-205.
- Wenden, A. L. (1998) Metacognitive Knowledge and Language Learning. *Applied Linguistics*, 19(4), 515-537.
- Wenden, A. L. & Rubin, J. (Eds.) (1987) *Learner Strategies in Language Learning*. Englewood Cliffs, NJ: Prentice Hall International.
- Wertsch, J. V. (1991) *Voices of the Mind: A Sociocultural Approach to Mediated Action*. Cambridge, Mass: Harvard University Press.
- Wertsch, J. V. (1998) *Mind as Action*. Oxford: Oxford University Press.
- Wertsch, J. V. (2007) Mediation. IN DANIELS, H., COLE, M. & WERTSCH, J. V. (Eds.) *The Cambridge Companion to Vygotsky*. Cambridge: Cambridge University Press.

- Wesche, M. & Paribakht, T. S. (1996) Assessing Second Language Vocabulary Knowledge: Depth Versus Breadth. *Canadian Modern Language Review*, 53(1), 13-40.
- Whistle, J. (1999) Concordancing with Students Using an 'Off-the-Web' Corpus. *ReCALL*, 11(2), 74-80.
- Wichmann, A., Fligelstone, S., McEnery, T. & Knowles, G. (Eds.) (1997) *Teaching and Language Corpora*, London; New York: Longman.
- Widdowson, H. G. (1978) *Teaching Language as Communication*. Oxford: Oxford University Press.
- Widdowson, H. G. (1984) Comments on the Implication of Interlanguage for Language Teaching. IN DAVIES, A., CRIPER, C. & HOWATT, A. P. R. (Eds.) *Interlanguage*. Edinburgh: Edinburgh University Press.
- Widdowson, H. G. (2000) On the Limitations of Linguistics Applied. *Applied Linguistics*, 21(1), 3-25.
- Wong-Fillmore, C. J. (1979) On Fluency. IN FILLMORE, C. J., KEMPLER, D. & WANG, W. S. Y. (Eds.) *Individual Differences in Language Ability and Behavior*. New York: Academic Press.
- Woodfield, H. (2010) What Lies Beneath?: Verbal Report in Interlanguage Requests in English. *Multilingua - Journal of Cross-Cultural and Interlanguage Communication*, 29(1), 1-27.
- Wray, A. (2000) Formulaic Sequences in Second Language Teaching: Principle and Practice. *Applied Linguistics*, 21(4), 463-489.
- Wray, A. (2002) *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press.
- Xiao, R. & Mcenery, T. (2006) Collocation, Semantic Prosody, and near Synonymy: A Cross-Linguistic Perspective. *Applied Linguistics*, 27(1), 103-129.
- Yeh, Y., Liou, H. C. & Li, Y. H. (2007) Online Synonym Materials and Concordancing for EFL College Writing. *Computer Assisted Language Learning*, 20(2), 131-152.
- Yoon, H. & Hirvela, A. (2004) ESL Student Attitudes toward Corpus Use in L2 Writing. *Journal of Second Language Writing*, 13, 257-283.
- Yorio, C. A. (1980) Conventionalized Language Forms and the Development of Communicative Competence. *TESOL Quarterly*, 14(4), 433-442.
- Young, K. A. (2005) Direct from the Source: The Value of 'Think-Aloud' Data in Understanding Learning. *Journal of Educational Enquiry*, 6(1), 19-33.
- Zechmeister, E. B. & Nyberg, S. E. (1982) *Human Memory: An Introduction to Research and Theory*. Monterey, CA: Brooks/Cole.
- Zhang, X. L. (1993) *English Collocations and Their Effect on the Writing of Native and Non-Native College Freshmen*. Unpublished doctoral dissertation. Indiana University of Pennsylvania.
- Zughoul, M. R. & Abdul-Fattah, H. S. (2003) Translational Collocational Strategies of Arab Learners of English. *Babel*, 49(1), 59-81.

## Dictionaries

- Collins COBUILD English Dictionary* (1995) 2<sup>nd</sup> edition, London: Harper Collins.
- Macmillan English Dictionary: For Advanced Learners of American English* (2004) Oxford: Palgrave Macmillan.
- Oxford Advanced Learner's Dictionary* (2007) 7<sup>th</sup> edition, Oxford: Oxford University Press.

*Oxford Collocations Dictionary for Students of English* (2002) Oxford: Oxford University Press.  
*The BBI Dictionary of English Word Combinations* (1997) Amsterdam: John Benjamins.



## Appendices

### Appendix 1: Consent form for the experimental group

Dear Student,

My name is Kuei-Ju Tsai. I am a doctoral student at the University of Bristol. The research is part of my PhD study.

This study aims to examine the effectiveness of using web-based concordancers to assist collocation learning, to explore underlying learning processes and to understand the learner perceptions of such a pedagogical approach.

If you agree to participate, you will need to take 5 learning sessions, a pretest and a post-test. You also need to provide 2 assignments of your writing class and complete a questionnaire after the learning sessions. The results of the study will be used for my research only. Any information you provide will be kept confidential, and **WILL NOT AFFECT YOUR GRADES AT SCHOOL**. You may refuse to participate or may withdraw at any time in the course of the study.

Please feel free to contact me if you have any queries about this study or the results:  
[kt6568@bristol.ac.uk](mailto:kt6568@bristol.ac.uk)

Please sign your name and date in the space below if you agree to participate in this study.

_____ Participant	_____ Date
_____ Researcher	_____ Date

**Appendix 2: Consent form for the control group**

Dear Student,

My name is Kuei-Ju Tsai. I am a doctoral student at the University of Bristol. The research is part of my PhD study.

This study aims to understand the collocational knowledge of English majors. If you agree to participate, you will need to take 2 collocation tests and provide 2 assignments of your writing class. The results of the study will be used for my research only. Any information you provide will be kept confidential, and **WILL NOT AFFECT YOUR GRADES AT SCHOOL**. You may refuse to participate or may withdraw at any time in the course of the study.

Please feel free to contact me if you have any queries about this study or the results:  
[kt6568@bristol.ac.uk](mailto:kt6568@bristol.ac.uk)

Please sign your name and date in the space below if you agree to participate in this study.

\_\_\_\_\_  
Participant

\_\_\_\_\_  
Date

\_\_\_\_\_  
Researcher

\_\_\_\_\_  
Date



### Appendix 3: Consent form for the think-aloud participants

Dear Student,

My name is Kuei-Ju Tsai. I am a doctoral student at the University of Bristol. The research is part of my PhD study. This study aims to examine the effectiveness of using web-based concordancers to assist collocation learning, to explore underlying learning processes and to understand the learner perceptions of such a pedagogical approach. This part of the study uses the think-aloud method to elicit the respondent's strategy use in the course of generalizing collocations from corpus data.

If you agree to participate, you will need to do a worksheet with the aid of web-based concordancers. At the same time, you need to think aloud: that is, verbalize your thinking processes. Your verbal reports will be audio-recorded. The results of the study will be used for my research only. Any information you provide will be kept confidential, and WILL NOT AFFECT YOUR GRADES AT SCHOOL. You may refuse to participate or may withdraw at any time in the course of the study.

Please feel free to contact me if you have any queries about this study or the results: [kt6568@bristol.ac.uk](mailto:kt6568@bristol.ac.uk)

Please sign your name and date in the space below if you agree to participate in this study.

_____	_____
Participant	Date
_____	_____
Researcher	Date

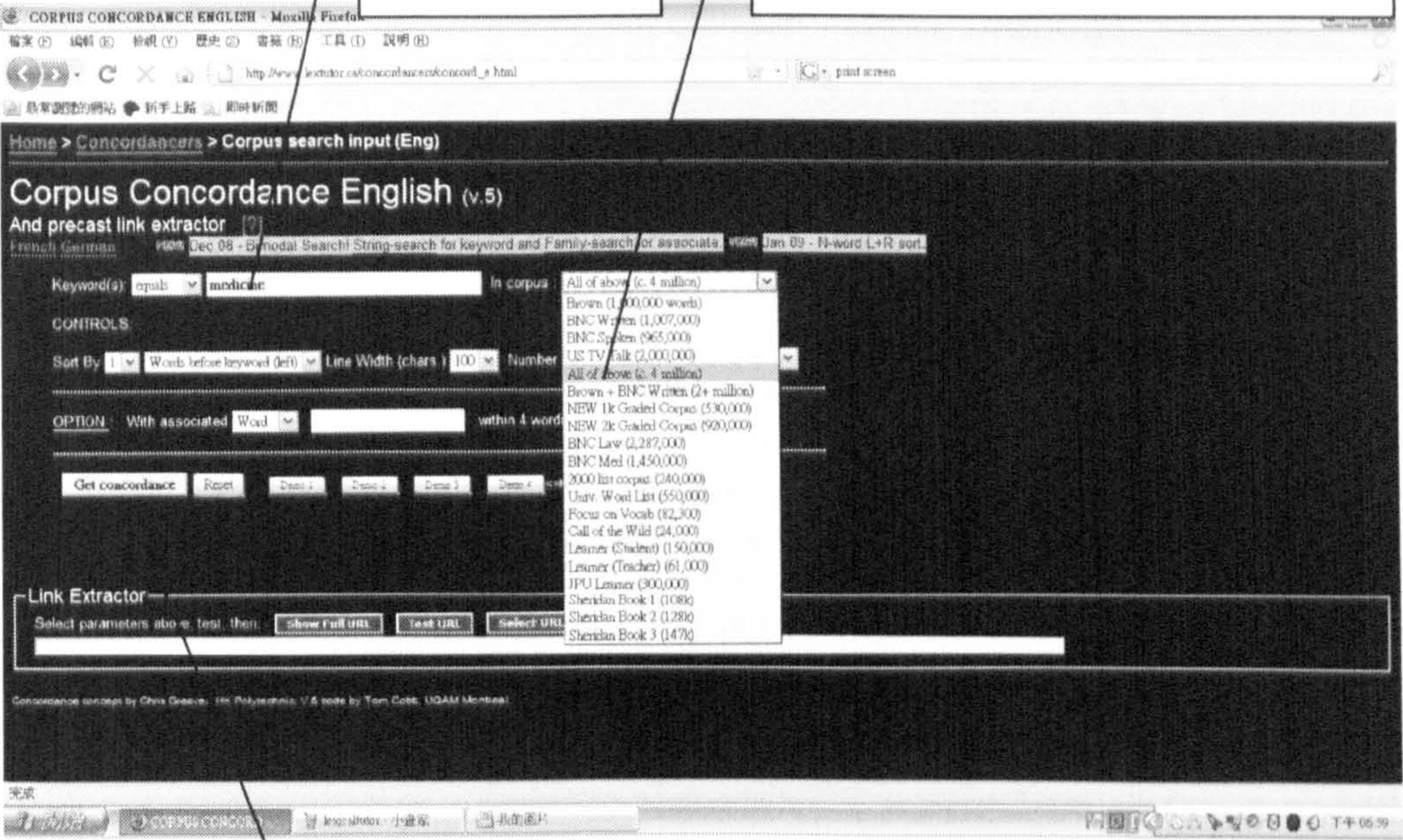


Appendix 4: The handout for DALC training session

Lexical Tutor ([http://www.lextutor.ca/concordancers/concord\\_e.html](http://www.lextutor.ca/concordancers/concord_e.html))

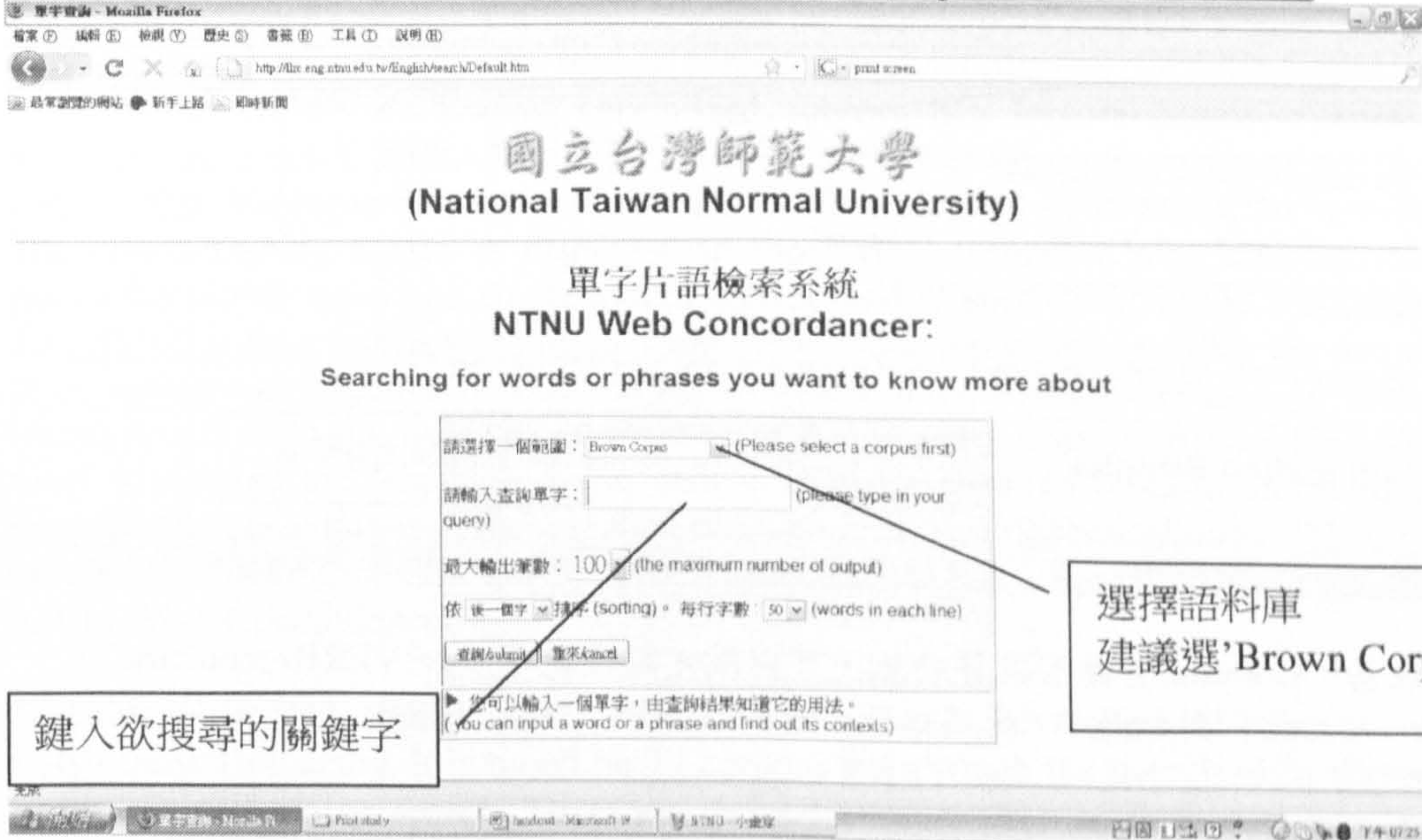
Step 1:  
鍵入關鍵字

Step2: 選擇語料庫，建議選第四個'All of above'，語料較多。



Step 3: 按下黃色鍵即開始搜尋

NTNU Web Concordancer (<http://nrc.eng.ntnu.edu.tw/english/search/Default.htm>)



單字片語檢索系統  
NTNU Web Concordancer:

Searching for words or phrases you want to know more about

請選擇一個範圍: Brown Corpus (Please select a corpus first)

請輸入查詢單字: (please type in your query)

最大輸出筆數: 100 (the maximum number of output)

依 後一個字 排序 (sorting) 每行字數: 50 (words in each line)

查詢/submit 重來/cancel

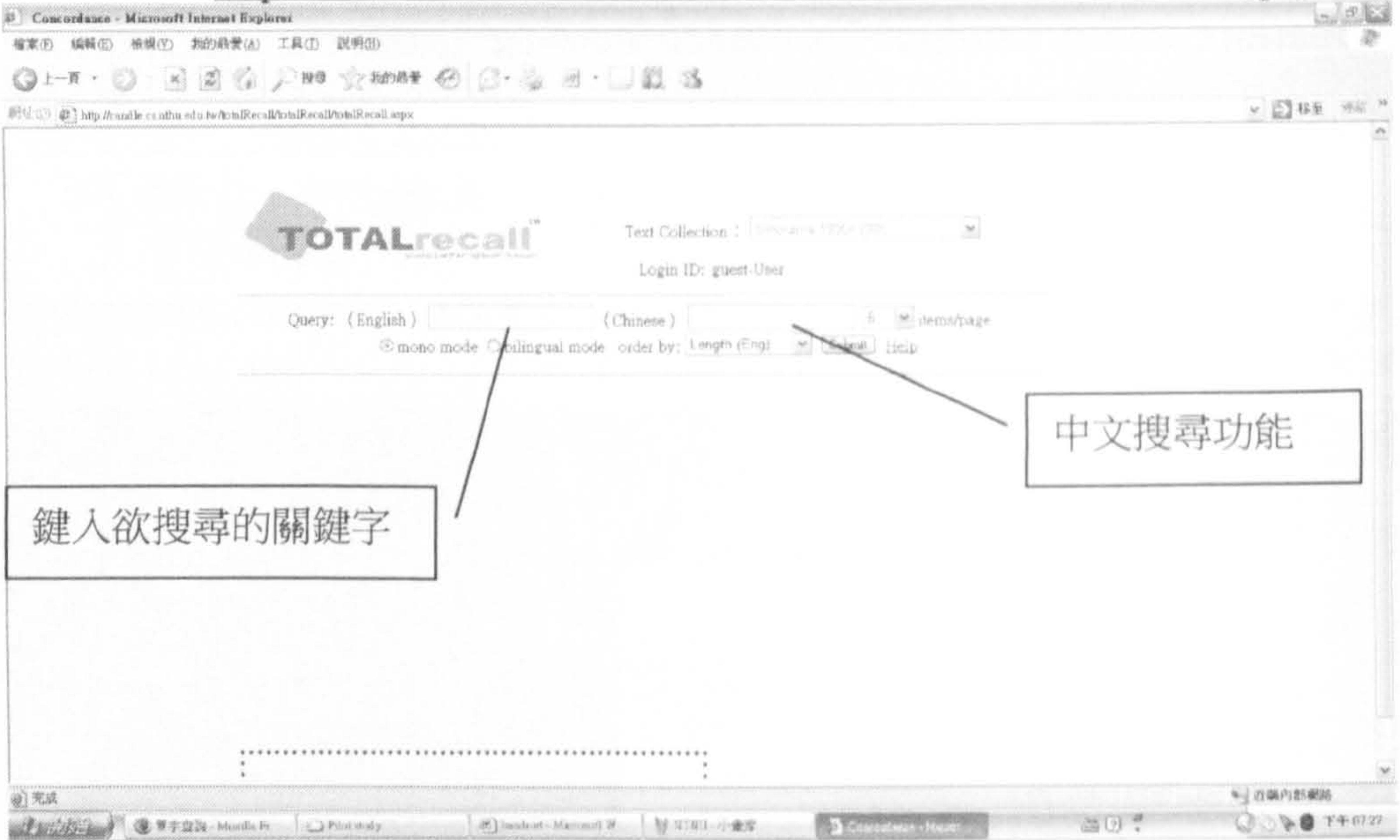
您可以輸入一個單字，由查詢結果知道它的用法。  
(you can input a word or a phrase and find out its contexts)

鍵入欲搜尋的關鍵字

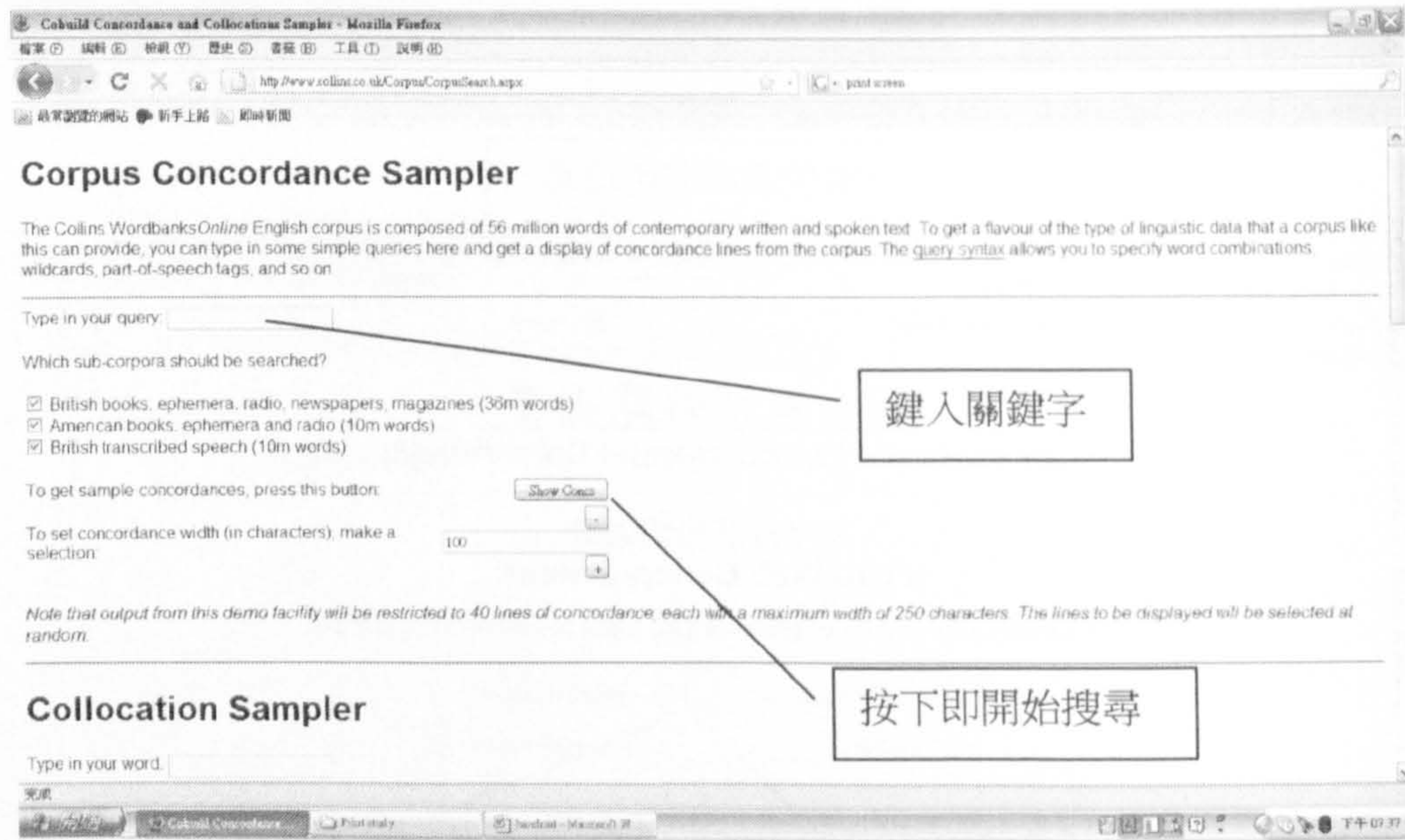
選擇語料庫  
建議選'Brown Corpus'



**TotalRecall** (<http://candle.cs.nthu.edu.tw/totalRecall/totalRecall/totalRecall.aspx>)



**Collins WordbanksOnline** (<http://www.collins.co.uk/Corpus/CorpusSearch.aspx>)



注意：Cobuild 有進階搜尋功能，可以限定詞性搜尋，如 'VERB+medicine'  
或 'JJ+wallet'，代碼如下：名詞 NOUN; 動詞 VERB; 形容詞 JJ; 副詞 RB



## Appendix 5: The designated web-based concordancers

### 1. TOTALrecall

TOTALrecall is a web-based concordancer developed by National Tsing Hwa University (NTHU) in Taiwan, the corpus of which is compiled from the articles of a Chinese-English bilingual magazine published in Taiwan - *Sinorama Magazine*. The magazine covers a wide range of topics about local life in Taiwan, including arts and culture, education, environment, finance and economy, health, science and technology, sports, and so on. The articles collected date from 1990 to 2000, containing over 2 million running words. Each article has both Chinese and English versions, which enables the readers to cross-check between two languages. TOTALrecall uses Chinese interface and supports Chinese-English bilingual search, allowing the user to submit queries in either language and obtain results displayed in both languages. The concordancing output is displayed in three columns with the English concordancing lines in the left column, the Chinese concordancing lines in the middle and the source of the concordancing lines in the right column, which trace back to the articles in which the concordancing lines are embedded. In addition to its bilingual search function, TOTALrecall offers another user-friendly function which automatically expands the scope of search to include not only the query word but also its inflected forms and singular/plural forms. As novice users of concordancers may not be flexible enough in their search to inflect verb forms to yield more results, the system provides more than needed by including variations of the query word.

### 2. NTNU Web Concordancer

The NTNU web concordancer was developed by National Taiwan Normal University. This concordancer also has Chinese interface but the corpora and the output are in English. There are a range of corpora behind the concordancer, including general corpora like the Brown corpus, specialized corpora like learner corpus, master's dissertations, articles from China News and TESOL Quarterly, transcripts of spoken corpora like Michigan Corpus of Academic Spoken English (MICASE), and so on. The concordancing output is displayed in the KWIC format. KWIC concordance places the search word in a single line of context and aligned vertically in a column. The NTNU web concordancer provides the sorting function which specifies the output to be arranged alphabetically based on the word left to the search word or right to the search word, or in the order of occurrence in the corpora. This function allows the users to arrange the output in a way that is easy for them to observe the word(s) immediately preceding or following the search word in an alphabetical order.

### 3. VLC Web Concordancer

The VLC (virtual language center) web concordancer is developed by Hong Kong Polytechnic University. It is based on 27 corpora from which the user need to choose one when performing search. There are general corpora such as Brown corpus and LOB corpus, as well as specialized corpora, such as literary texts like *Lord of the Rings*, religious texts like the Bible or the Koran, Hong Kong government documents, student writings, articles of the *TIMES* magazine, and so on (see Table 1 for the complete range of corpora and their word count), with the size ranging from 56,336



words to over 3 millions words. The VLC web concordancer provides monolingual concordance in four languages – English, Chinese, French and Japanese in addition to Chinese-English bilingual concordance. However, the parallel texts used for bilingual concordance are predominantly literary texts, such as *Alice in Wonderland*, *Adventures of Huckleberry Finn*, or Chinese fables and novels. The bilingual concordance would not be recommended to the participants in the present study as the study focuses attention to the collocations used in general corpus data instead of that in specialized corpora. The web concordancer displays output in KWIC (keyword in context) mode and allows the users to explore the context in which the query word is embedded with a click on the word of that particular concordance line. In addition to KWIC, the VLC web concordancer offers another option of output mode by displaying output in complete sentences. The website also has online dictionary which enables the users to check word meanings if they encounter unknown words as they concordance.

#### 4. The Collins WordbanksOnline

The Collins WordbanksOnline English corpus is composed of 56 million words of contemporary written and spoken text. The corpus is drawn from the Bank of English which contains over 450 million words from a wide range of written and spoken sources dated from 1990. The Bank of English was launched for the purpose of dictionary compilation and language study by COBUILD (a division of HarperCollins Publishers) and the University of Birmingham. The Collins WordbanksOnline English corpus allows the users to investigate from a number of subcorpora, including 36 million words from British books, ephemera, radio, newspapers, magazines; 10 million words from American books, ephemera and radio; and 10 million words from British transcribed speech. WordbanksOnline is a monolingual concordancer, supporting search function in English only. The corpus data is tagged and thus provides part-of-speech queries that are more sophisticated than the simple query function. For instance, to search for the verb collocates of the noun *basis*, one can perform a simple query by entering 'basis' to see what collocates come before and after the query word *basis* and manually screen the output for verb collocates. However, with a concordancer like WordbanksOnline that supports part-of-speech queries, s/he can limit the options of collocates to verbs only by specifying the part-of-speech of the query and entering commands such as 'VERB+the+basis'. The concordance output is displayed in the format of KWIC. The weakness of this program is that the output line does not trace back to the text in which it was embedded. That is, if a user reads a concordance line and wishes to investigate further the immediate or wider context where it has occurred, s/he would not have access to the context as the program does not display concordance lines along with the texts from which they are originated.

## Appendix 6: The pilot test

### I. Choose one INCORRECT option

For example: She \_\_\_\_\_ notes of the lecture.

(A) takes (B) keeps (C) does

Both *take notes* and *keep notes* are correct collocations, but *\*do notes* is not. Therefore, (C) is the incorrect option.

- \_\_\_\_\_ 1. We need to \_\_\_\_\_ a plan before we can start on the project.  
(A) take (B) devise (C) make
- \_\_\_\_\_ 2. The writer \_\_\_\_\_ inspiration from his travels around the world.  
(A) draws (B) generates (C) takes
- \_\_\_\_\_ 3. The professor \_\_\_\_\_ a speech about current economic situations.  
(A) sent (B) delivered (C) gave
- \_\_\_\_\_ 4. Good visuals and diagrams are the magazine's most \_\_\_\_\_ features.  
(A) distinctive (B) differentiating (C) distinguishing
- \_\_\_\_\_ 5. Support for the plan remained \_\_\_\_\_ solid.  
(A) rock (B) absolutely (C) hardly
- \_\_\_\_\_ 6. Everybody rushed to a shelter when the siren \_\_\_\_\_.  
(A) went (B) called (C) sounded
- \_\_\_\_\_ 7. The students are asked to \_\_\_\_\_ a journal in which they summarize the daily readings.  
(A) keep (B) write (C) take
- \_\_\_\_\_ 8. After two years in the Army, he \_\_\_\_\_ a job at The New York Times.  
(A) landed (B) got (C) made
- \_\_\_\_\_ 9. I had always intended to pay him a visit, but my busy career made it hard for me to \_\_\_\_\_ a trip outside L.A.  
(A) set (B) organize (C) plan
- \_\_\_\_\_ 10. A \_\_\_\_\_ determination will get you through the obstacles as they arise.  
(A) dogged (B) powerful (C) fierce
- \_\_\_\_\_ 11. Hummingbirds \_\_\_\_\_ a journey of more than 6,000 kilometers round trip from their breeding grounds in autumn to their return the following spring.  
(A) make (B) undertake (C) fly
- \_\_\_\_\_ 12. In early July, Mr. Farley and other executives announced that something had gone terribly wrong and that profits would not \_\_\_\_\_ expectations.  
(A) achieve (B) fulfill (C) meet

- \_\_\_\_\_13. Feeling nervous and growing cold as the temperature \_\_\_\_\_ rapidly, his arms and legs trembled.  
(A) dropped (B) fell (C) reduced
- \_\_\_\_\_14. Mr. Taylor earned an Avery Fisher Career Grant, and last year he \_\_\_\_\_ an award from the American Pianists Association.  
(A) took (B) received (C) won
- \_\_\_\_\_15. The soccer player is not \_\_\_\_\_ injured and will play in Friday's game against Arizona.  
(A) seriously (B) heavily (C) badly
- \_\_\_\_\_16. It is estimated that only 20 percent of US workers currently have the level of skills and training to \_\_\_\_\_ the challenge of international competition.  
(A) receive (B) face (C) meet
- \_\_\_\_\_17. I was taught the Zen (禪) concept of happiness, which was to \_\_\_\_\_ satisfaction in small things, such as the fresh air in the morning.  
(A) find (B) reach (C) take
- \_\_\_\_\_18. Finally, he \_\_\_\_\_ the courage to tell his parents that he had dropped out of school.  
(A) kept up (B) plucked up (C) worked up
- \_\_\_\_\_19. If you're going to New York - and you want to take both your heart and your wallet home - \_\_\_\_\_ alert when sightseeing there.  
(A) stay (B) hold (C) remain
- \_\_\_\_\_20. Many churches have icons showing saints. Icons on the walls of the church have \_\_\_\_\_ significance.  
(A) great (B) considerable (C) extensive
- \_\_\_\_\_21. The poet gained a \_\_\_\_\_ audience through his use of simple language about everyday subjects, many of them rooted in his childhood memories.  
(A) large (B) many (C) wide
- \_\_\_\_\_22. Mars is \_\_\_\_\_ visible in the south and southwest during June and July evenings, but has faded greatly from its peak brilliance back in March.  
(A) obviously (B) clearly (C) plainly
- \_\_\_\_\_23. It took them three years to \_\_\_\_\_ approval to build a housing project for the elderly.  
(A) acquire (B) gain (C) win



## II. Fill in the blank

1. The meeting aims to address a w (adj.) range of issues, from the environment to the educational system.
2. I don't usually lend people money, but in your case I'll m (v.) an exception.
3. Nobody had paid much attention, including Steven. It was only when the man climbed up onto a table and started singing a birthday song that Steven began to t (v.) notice.
4. When people become dependent on various machines and abandon other means of transportation in favor of the automobile, a change o (v.) in thinking, values, and life style.
5. He has p (adj.) hearing as a result of ear infections (感染).
6. The university needed to s (v.) a balance between the conflicting interests of athletics and academics, guaranteeing players a good education at the expense of winning teams.
7. Cultures can hardly be understood only if one does not know the language of the culture. Furthermore, in my own view, one can not f (adv.) appreciate one's own culture unless one can also see it through the lens of another culture.
8. Even though separating the twins meant one would die and the operation would cost more than a half a million dollars, their parents and doctors said it was w (adj.) a try.
9. The girl was so thin and pale (苍白) that she seemed to have just recovered (復原) from a s (adj.) illness: her large dark eyes shone much brighter on her pale face.
10. In 1886, the Coca Cola Company was developed but it wasn't until 1898 that the f (adj.) competitor Pepsi-Cola entered into the market. These 2 companies are the two major players that dominate the soft-drink industry.
11. Scientists m (v.) an interpretation based on the data available.
12. Last year, only 15 of the city's 2,500 teachers were rated unsatisfactory, while the v (adj.) majority were rated either superior or above average.
13. Local people have e (v.) their dismay at the cuts in social services.
14. The pain in Tom's right hand was a c (adj.) reminder of that fight.
15. The singer discovered that nothing gave him greater pleasure than hearing an audience b (v.) into applause after he sang.
16. We agreed to do the work free of charge as a g (n.) of goodwill.
17. Disaster s (v.) as the heavy rain flooded the village and claimed 30 lives.
18. The president m (v.) an announcement that an agreement for peace had been made in Iraq (伊拉克).
19. She went out to the garden for a b (n.) of fresh air.
20. By 1978, Levis was already a household n (n.) everywhere in the world. This year, Levi Strauss has annual revenue (營收) of \$ 6.1 billion.
21. The song always makes her c (v.) a smile and shed a tear at the same time.
22. For men who make their living playing football, head trauma (創傷) is an o (adj.) hazard as there is a risk of hurting their heads during work.
23. The doctor m (v.) a diagnosis and developed a plan for treatment for the patient.

24. Although many students who attend community colleges do so for economic and convenience reasons, some students attend community colleges because their high school grades or college admittance test scores are too low to g\_\_\_\_\_ (v.) admittance to their university of choice.
25. The government official was s\_\_\_\_\_ (adv.) reprimanded for illegally accepting the money.
26. There appears to be considerable doubts on the part of the teachers to f\_\_\_\_\_ (adv.) embrace the concept of bringing students with special needs into their classrooms. In theory, they support the concept, but in practice they are still uncertain and want to retain special education classrooms.
27. Scientists have m\_\_\_\_\_ (v.) an observation that could revolutionize thinking about evolution (演化) and even cancer. They report that genetic mutations (基因突變) build up in cells at a faster rate than anyone thought.

## Appendix 7: The pretest for the main study

### I. Choose one INCORRECT option

For example: She \_\_\_\_\_ notes of the lecture.

(A) takes (B) keeps (C) does

Both *take notes* and *keep notes* are correct collocations, but *\*do notes* is not. Therefore, (C) is the incorrect option.

- \_\_\_\_\_ 1. We need to \_\_\_\_\_ a plan before we can start on the project.  
(A) take (B) devise (C) make
- \_\_\_\_\_ 2. The writer \_\_\_\_\_ inspiration from his travels around the world.  
(A) draws (B) generates (C) takes
- \_\_\_\_\_ 3. The professor \_\_\_\_\_ a speech about current economic situations.  
(A) sent (B) delivered (C) gave
- \_\_\_\_\_ 4. Good visuals and diagrams are the magazine's most \_\_\_\_\_ features.  
(A) distinctive (B) differentiating (C) distinguishing
- \_\_\_\_\_ 5. Support for the plan remained \_\_\_\_\_ solid.  
(A) rock (B) absolutely (C) hardly
- \_\_\_\_\_ 6. Everybody rushed to a shelter when the siren \_\_\_\_\_.  
(A) went (B) called (C) sounded
- \_\_\_\_\_ 7. The students are asked to \_\_\_\_\_ a journal in which they summarize the daily readings.  
(A) keep (B) write (C) take
- \_\_\_\_\_ 8. After two years in the Army, he \_\_\_\_\_ a job at The New York Times.  
(A) landed (B) got (C) made
- \_\_\_\_\_ 9. I had always intended to pay him a visit, but my busy career made it hard for me to \_\_\_\_\_ a trip outside L.A.  
(A) set (B) organize (C) plan
- \_\_\_\_\_ 10. A \_\_\_\_\_ determination will get you through the obstacles as they arise.  
(A) dogged (B) powerful (C) fierce
- \_\_\_\_\_ 11. Hummingbirds \_\_\_\_\_ a journey of more than 6,000 kilometers round trip from their breeding grounds in autumn to their return the following spring.  
(A) make (B) undertake (C) fly
- \_\_\_\_\_ 12. In early July, Mr. Farley and other executives announced that something had gone terribly wrong and that profits would not \_\_\_\_\_ expectations.  
(A) achieve (B) fulfill (C) meet
- \_\_\_\_\_ 13. Feeling nervous and growing cold as the temperature \_\_\_\_\_ rapidly, his arms and legs trembled.  
(A) dropped (B) fell (C) reduced
- \_\_\_\_\_ 14. Mr. Taylor earned an Avery Fisher Career Grant, and last year he \_\_\_\_\_ an award from the American Pianists Association.  
(A) took (B) received (C) won
- \_\_\_\_\_ 15. The soccer player is not \_\_\_\_\_ injured and will play in Friday's game against Arizona.  
(A) seriously (B) heavily (C) badly
- \_\_\_\_\_ 16. It is estimated that only 20 percent of US workers currently have the level of skills and training to \_\_\_\_\_ the challenge of international competition.  
(A) receive (B) face (C) meet



- \_\_\_\_\_17. I was taught the Zen (禪) concept of happiness, which was to \_\_\_\_\_  
satisfaction in small things, such as the fresh air in the morning.  
(A) find (B) reach (C) take
- \_\_\_\_\_18. Finally, he \_\_\_\_\_ the courage to tell his parents that he had dropped  
out of school.  
(A) kept up (B) plucked up (C) worked up
- \_\_\_\_\_19. If you're going to New York - and you want to take both your heart and your  
wallet home - \_\_\_\_\_ alert when sightseeing there.  
(A) stay (B) hold (C) remain
- \_\_\_\_\_20. Many churches have icons showing saints. Icons on the walls of the church  
have \_\_\_\_\_ significance.  
(A) great (B) considerable (C) extensive
- \_\_\_\_\_21. The poet gained a \_\_\_\_\_ audience through his use of simple language  
about everyday subjects, many of them rooted in his childhood memories.  
(A) large (B) many (C) wide
- \_\_\_\_\_22. Mars (火星) is \_\_\_\_\_ visible in the south and southwest during June  
and July evenings, but has faded greatly from its peak brilliance back in  
March.  
(A) obviously (B) clearly (C) plainly
- \_\_\_\_\_23. It took them three years to \_\_\_\_\_ approval to build a housing project  
for the elderly.  
(A) reach (B) gain (C) win

## II. Fill in the blank

1. The meeting aims to address a w\_\_\_\_\_ (adj.) range of issues, from the environment to the educational system.
2. I don't usually lend people money, but in your case I'll m\_\_\_\_\_ (v.) an exception.
3. Nobody had paid much attention, including Steven. It was only when the man climbed up onto a table and started singing a birthday song that Steven began to t\_\_\_\_\_ (v.) notice.
4. When people become dependent on various machines and abandon other means of transportation in favor of the automobile, a change o\_\_\_\_\_ (v.) in thinking, values, and life style.
5. He has p\_\_\_\_\_ (adj.) hearing as a result of ear infections (感染).
6. The university needed to s\_\_\_\_\_ (v.) a balance between the conflicting interests of athletics and academics, guaranteeing players a good education at the expense of winning teams.

7. Cultures can hardly be understood only if one does not know the language of the culture. Furthermore, one can not f (adv.) appreciate one's own culture unless one can also see it through the lens of another culture.
8. Even though separating the twins meant one would die and the operation would cost more than a half a million dollars, their parents and doctors said it was w (adj.) a try.
9. The girl was so thin and pale (蒼白) that she seemed to have just recovered (復原) from a s (adj.) illness: her large dark eyes shone much brighter on her pale face.
10. In 1886, the Coca Cola Company was developed but it wasn't until 1898 that the f (adj.) competitor Pepsi-Cola entered into the market. These 2 companies are the two major players that dominate the soft-drink industry.
11. The artist's creativity comes from his k (adj.) observation on life and nature.
12. Disaster s (v.) as the heavy rain flooded the village and claimed 30 lives.
13. Last year, only 15 of the city's 2,500 teachers were rated unsatisfactory, while the v (adj.) majority were rated either superior or above average.
14. We agreed to do the work free of charge as a g (n.) of goodwill.
15. Local people have e (v.) their dismay at the cuts in social services.
16. The pain in Tom's right hand was a c (adj.) reminder of that fight.
17. The singer discovered that nothing gave him greater pleasure than hearing an audience b (v.) into applause after he sang.
18. The president m (v.) an announcement that an agreement for peace had been made in Iraq (伊拉克).
19. She went out to the garden for a b (n.) of fresh air.
20. By 1978, Levis was already a household n (n.) everywhere in the world. This year, Levi Strauss has annual revenue (營收) of \$ 6.1 billion.
21. Climbing the mountain gave him a great s (n.) of achievement.
22. For men who make their living playing football, head trauma (創傷) is an o (adj.) hazard as there is a risk of hurting their heads during work.



23. Coming up with ideas to paint has never been a problem for me because I've always had a v\_\_\_\_\_ (adj.) imagination.
24. Although many students who attend community colleges do so for economic and convenience reasons, some students attend community colleges because their high school grades or college admittance test scores are too low to g\_\_\_\_\_ (v.) admittance to their university of choice.
25. The government official was s\_\_\_\_\_ (adv.) reprimanded for illegally accepting the money.
26. Her d\_\_\_\_\_ (adj.) devotion and love for her country and for her people - her nobility, strength, and courage -- are all qualities I aspire to (渴求) as a human being.
27. There appears to be considerable doubts on the part of the teachers to f\_\_\_\_\_ (adv.) embrace the concept of bringing students with special needs into their classrooms. In theory, they support the concept, but in practice they are still uncertain and want to retain special education classrooms.

## Appendix 8: The post-test for the main study

### I. Choose one CORRECT option

- \_\_\_\_\_ 1. We need to \_\_\_\_\_ a plan before we can start on the project.  
(A) do (B) make (C) take
- \_\_\_\_\_ 2. The writer \_\_\_\_\_ inspiration from his travels around the world.  
(A) brings (B) draws (C) generates
- \_\_\_\_\_ 3. The professor \_\_\_\_\_ a speech about current economic situations.  
(A) delivered (B) sent (C) talked
- \_\_\_\_\_ 4. Good visuals are the magazine's most \_\_\_\_\_ features.  
(A) distinctive (B) differing (C) discriminating
- \_\_\_\_\_ 5. Support for the plan remained \_\_\_\_\_ solid.  
(A) hard (B) firm (C) rock
- \_\_\_\_\_ 6. Everybody rushed to a shelter (避難所) when the siren \_\_\_\_\_.  
(A) call (B) shouted (C) went
- \_\_\_\_\_ 7. The students are asked to \_\_\_\_\_ a journal in which they summarize the daily readings.  
(A) do (B) keep (C) take
- \_\_\_\_\_ 8. After two years in the Army, he \_\_\_\_\_ a job at The New York Times.  
(A) landed (B) made (C) offered
- \_\_\_\_\_ 9. I had always wanted to visit him, but my busy career made it hard for me to \_\_\_\_\_ a trip outside L.A.  
(A) set (B) manage (C) plan
- \_\_\_\_\_ 10. A \_\_\_\_\_ determination will get you through the obstacles (障礙) as they arise.  
(A) dogged (B) powerful (C) tough
- \_\_\_\_\_ 11. Hummingbirds (蜂鳥) \_\_\_\_\_ a journey of more than 6,000 km round trip from their breeding grounds in autumn to their return the following spring.  
(A) fly (B) undertake (C) step
- \_\_\_\_\_ 12. In early July, Mr. Farley and other executives announced that something had gone terribly wrong and that profits would not \_\_\_\_\_ expectations.  
(A) achieve (B) attain (C) meet
- \_\_\_\_\_ 13. Feeling nervous and growing cold as the temperature \_\_\_\_\_ rapidly, his arms and legs trembled.  
(A) cut (B) dropped (C) reduced

- \_\_\_\_\_14. The soccer player is not \_\_\_\_\_ injured and will play in Friday's game.  
(A) badly (B) heavily (C) strongly
- \_\_\_\_\_15. Mr. Taylor \_\_\_\_\_ an award from the American Pianists Association.  
(A) took (B) gained (C) won
- \_\_\_\_\_16. It is estimated that only 20 percent of US workers have the level of skills and training to \_\_\_\_\_ the challenge of international competition.  
(A) meet (B) receive (C) take
- \_\_\_\_\_17. I was taught the Zen (禪) concept of happiness, which was to \_\_\_\_\_ satisfaction in small things, such as the fresh air in the morning.  
(A) draw (B) find (C) reach
- \_\_\_\_\_18. Finally, he \_\_\_\_\_ the courage to tell his parents that he had dropped out of school.  
(A) kept up (B) caught up (C) worked up
- \_\_\_\_\_19. Many churches have icons (聖像) showing saints. Icons on the walls of the church have \_\_\_\_\_ significance.  
(A) considerable (B) extensive (C) large
- \_\_\_\_\_20. The poet gained a \_\_\_\_\_ audience through his use of simple language about everyday subjects.  
(A) extensive (B) many (C) wide
- \_\_\_\_\_21. Mars (火星) is \_\_\_\_\_ visible in the south during July evenings.  
(A) apparently (B) obviously (C) plainly
- \_\_\_\_\_22. It took them three years to \_\_\_\_\_ approval to build a housing project for the elderly.  
(A) acquire (B) gain (C) reach

## II. Fill in the blank

- I don't usually lend people money, but in your case I'll m \_\_\_\_\_ (v.) an exception.
- Nobody had paid much attention, including Steven. It was only when the man climbed up onto a table and started singing a birthday song that Steven began to t \_\_\_\_\_ (v.) notice.
- The singer felt that nothing gave him greater pleasure than hearing an audience b \_\_\_\_\_ (v.) into applause after he sang.



4. The university needed to s\_\_\_\_\_ (v.) a balance between the conflicting interests of athletics and academics, guaranteeing players a good education at the expense of winning teams.
5. Local people have e\_\_\_\_\_ (v.) their dismay at the cuts in social services.
6. The president m\_\_\_\_\_ (v.) an announcement that an agreement for peace had been made in Iraq (伊拉克).
7. Although many students who attend community colleges do so for economic and convenience reasons, some students attend community colleges because their college admittance test scores are too low to g\_\_\_\_\_ (v.) admittance to their university of choice.
8. Disaster s\_\_\_\_\_ (v.) as the heavy rain flooded the village and claimed 30 lives.
9. When people become dependent on various machines and abandon other means of transportation in favor of the automobile, a change o\_\_\_\_\_ (v.) in thinking, values, and life style.
10. The girl was so thin and pale (苍白) that she seemed to have just recovered (复原) from a s\_\_\_\_\_ (adj.) illness.
11. The meeting aims to address a w\_\_\_\_\_ (adj.) range of issues, from the environment to the educational system.
12. The pain in Tom's right hand was a c\_\_\_\_\_ (adj.) reminder of that fight.
13. He has p\_\_\_\_\_ (adj.) hearing as a result of ear infections (感染).
14. Thousands of police were put on f\_\_\_\_\_ (adj.) alert at all main roads leading to the city.
15. Even though separating the twins meant one would die and the operation would cost more than a half a million dollars, their parents and doctors said it was w\_\_\_\_\_ (adj.) a try.
16. In 1886, the Coca Cola Company was developed but it wasn't until 1898 that the f\_\_\_\_\_ (adj.) competitor Pepsi-Cola entered into the market. These 2 companies are the two major players that dominate the soft-drink industry.
17. The painter's creativity comes from his k\_\_\_\_\_ (adj.) observation on life and nature.

18. Last year, only 15 of the city's 2,500 teachers were rated unsatisfactory, while the v (adj.) majority were rated either superior or above average.
19. For men who make their living playing football, head trauma (創傷) is an o (adj.) hazard as there is a risk of hurting their heads during work.
20. Coming up with ideas to paint has never been a problem for me because I've always had a v (adj.) imagination.
21. Her d (adj.) devotion and love for her country and for her people - her strength and courage -- are the qualities I aspire to (渴求) as a human being.
22. We agreed to do the work free of charge as a g (n.) of goodwill.
23. She went out to the garden for a h (n.) of fresh air.
24. By 1978, Levis was already a household n (n.) everywhere in the world. This year, Levi Strauss has annual revenue (營收) of \$ 6.1 billion.
25. Climbing the mountain gave him a great s (n.) of achievement.
26. Cultures can hardly be understood only if one does not know the language of the culture. Furthermore, one can not f (adv.) appreciate one's own culture unless one can also see it through the lens of another culture.
27. The government official was s (adv.) reprimanded for illegally accepting the money.
28. There appears to be doubts on the part of the teachers to f (adv.) embrace the concept of bringing students with special needs into their classrooms. In theory, they support the concept, but in practice they are still uncertain and want to retain (保留) special education classrooms.

## Appendix 9: The worksheets for DALC intervention sessions

### Worksheet (1)

#### I. Choose one INCORRECT option

- \_\_\_\_\_ 1. I had always intended to pay him a visit, but my busy career made it hard for me to \_\_\_\_\_ a trip outside L.A.  
(A) set (B) organize (C) plan
- \_\_\_\_\_ 2. Hummingbirds (蜂鳥) \_\_\_\_\_ a journey of more than 6,000 km round trip from their breeding grounds in autumn to their return the following spring.  
(A) make (B) undertake (C) fly
- \_\_\_\_\_ 3. After two years in the Army, he \_\_\_\_\_ a job at The New York Times.  
(A) landed (B) got (C) made
- \_\_\_\_\_ 4. I was taught the Zen (禪) concept of happiness, which was to \_\_\_\_\_ satisfaction in small things, such as the fresh air in the morning.  
(A) find (B) reach (C) take
- \_\_\_\_\_ 5. It took them three years to \_\_\_\_\_ approval to build a housing project for the elderly.  
(A) acquire (B) gain (C) win

#### II. Fill in the blank

1. Nobody had paid much attention, including Steven. It was only when the man climbed up onto a table and started singing a birthday song that Steven began to t\_\_\_\_\_ (v.) notice.
2. Even though separating the twins meant one would die and the operation would cost more than a half a million dollars, their parents and doctors said it was w\_\_\_\_\_ (adj.) a try.
3. Coming up with ideas to paint has never been a problem for me because I've always had a v\_\_\_\_\_ (adj.) imagination.
4. We agreed to do the work free of charge as a g\_\_\_\_\_ (n.) of goodwill.
5. Cultures can hardly be understood only if one does not know the language of the culture. Furthermore, in my own view, one can not f\_\_\_\_\_ (adv.) appreciate one's own culture unless one can also see it through the lens of another culture.



## Worksheet (2)

### I. Choose one INCORRECT option

- \_\_\_\_\_ 1. The soccer player is not \_\_\_\_\_ injured and will play in Friday's game against Arizona.  
(A) badly (B) heavily (C) seriously (D) severely
- \_\_\_\_\_ 2. Feeling nervous and growing cold as the temperature \_\_\_\_\_ rapidly, his arms and legs trembled.  
(A) dropped (B) fell (C) plummeted (D) reduced
- \_\_\_\_\_ 3. It is estimated that only 20 percent of US workers currently have the level of skills and training to \_\_\_\_\_ the challenge of international competition.  
(A) accept (B) face (C) meet (D) receive
- \_\_\_\_\_ 4. A \_\_\_\_\_ determination will get you through the obstacles as they arise.  
(A) dogged (B) fierce (C) powerful (D) strong
- \_\_\_\_\_ 5. Finally, he \_\_\_\_\_ the courage to tell his parents that he had dropped out of school.  
(A) got up (B) kept up (C) plucked up (D) worked up

### II. Fill in the blank

1. Last year, only 15 of the city's 2,500 teachers were rated unsatisfactory, while the o\_\_\_\_\_ / s\_\_\_\_\_ / v\_\_\_\_\_ (adj.) majority were rated either superior or above average.
2. In 1886, the Coca Cola Company was developed but it wasn't until 1898 that the f\_\_\_\_\_ / s\_\_\_\_\_ (adj.) competitor Pepsi-Cola entered into the market. These companies are the two major players that dominate the soft-drink industry.
3. For men who make their living playing football, head trauma (創傷) is an o\_\_\_\_\_ (adj.) hazard as there is a risk of hurting their heads during work.
4. Disaster h\_\_\_\_\_ / o\_\_\_\_\_ / s\_\_\_\_\_ (v.) as the heavy rain flooded the village and claimed 30 lives.
5. When people become dependent on various machines and abandon other means of transportation in favor of the automobile, a change o\_\_\_\_\_ (v.) / t\_\_\_\_\_ (vp) in thinking, values, and life style.

## Worksheet (3)

### I. Choose one INCORRECT option

- \_\_\_\_\_ 1. We need to \_\_\_\_\_ a plan before we can start on the project.  
(A) develop (B) devise (C) make (D) take
- \_\_\_\_\_ 2. In early July, Mr. Farley and other executives announced that something had gone terribly wrong and that profits would not \_\_\_\_\_ expectations.  
(A) achieve (B) fulfill (C) meet (D) satisfy
- \_\_\_\_\_ 3. Many churches have icons showing saints. Icons on the walls of the church have \_\_\_\_\_ significance.  
(A) considerable (B) extensive (C) great (D) profound

### II. Fill in the blank

1. The students are asked to k\_\_\_\_\_ (v.) / w\_\_\_\_\_ (v.) a journal in which they summarize the daily readings.
2. I don't usually lend people money, but in your case I'll m\_\_\_\_\_ (v.) an exception.
3. Although many students who attend community colleges do so for economic and convenience reasons, some students attend community colleges because their high school grades or college admittance test scores are too low to g\_\_\_\_\_ (v.) admittance to their university of choice.
4. Climbing the mountain gave him a great s\_\_\_\_\_ (n.) of achievement.
5. The girl was so thin and pale (蒼白) that she seemed to have just recovered (復原) from a s\_\_\_\_\_ / s\_\_\_\_\_ (adj.) illness: her large dark eyes shone much brighter on her pale face.
6. The pain in Tom's right hand was a c\_\_\_\_\_ (adj.) reminder of that fight.
7. There appears to be considerable doubts on the part of the teachers to f\_\_\_\_\_ (adv.) embrace the concept of bringing students with special needs into their classrooms. In theory, they support the concept, but in practice they are still uncertain and want to retain special education classrooms.

## Worksheet (4)

### I. Choose one INCORRECT option

- \_\_\_\_\_ 1. The professor \_\_\_\_\_ a speech about current economic situations.  
(A) delivered (B) gave (C) made (D) talked
- \_\_\_\_\_ 2. The poet gained a \_\_\_\_\_ audience through his use of simple language about everyday subjects, many of them rooted in his childhood memories.  
(A) large (B) many (C) mass (D) wide
- \_\_\_\_\_ 3. Mr. Taylor earned an Avery Fisher Career Grant, and last year he \_\_\_\_\_ an award from the American Pianists Association.  
(A) got (B) received (C) took (D) won

### II. Fill in the blank

1. Thousands of police were put on f \_\_\_\_\_ (adj.) alert at all main roads leading to the city.
2. The president m \_\_\_\_\_ (v.) an announcement that an agreement for peace had been made in Iraq (伊拉克).
3. The singer discovered that nothing gave him greater pleasure than hearing an audience b \_\_\_\_\_ / b \_\_\_\_\_ (v.) into applause after he sang.
4. Local people have e \_\_\_\_\_ (v.) their dismay at the cuts in social services.
5. The government official was s \_\_\_\_\_ (adv.) reprimanded for illegally accepting the money.
6. The meeting aims to address a b \_\_\_\_\_ / w \_\_\_\_\_ (adj.) range of issues, from the environment to the educational system.
7. By 1978, Levis was already a household n \_\_\_\_\_ (n.) everywhere in the world. This year, Levi Strauss has annual revenue (营收) of \$ 6.1 billion.



## Worksheet (5)

### I. Choose one INCORRECT option

- \_\_\_\_\_ 1. Support for the plan remained \_\_\_\_\_ solid.  
(A) rock (B) absolutely (C) hardly
- \_\_\_\_\_ 2. Everybody rushed to a shelter when the siren \_\_\_\_\_.  
(A) went (B) called (C) sounded
- \_\_\_\_\_ 3. Good visuals and diagrams are the magazine's most \_\_\_\_\_ features.  
(A) distinctive (B) differentiating (C) distinguishing
- \_\_\_\_\_ 4. Mars (火星) is \_\_\_\_\_ visible in the south and southwest during June and July evenings, but has faded greatly from its peak brilliance back in March.  
(A) obviously (B) clearly (C) plainly
- \_\_\_\_\_ 5. The writer \_\_\_\_\_ inspiration from his travels around the world.  
(A) draws (B) generates (C) takes

### II. Fill in the blank

1. The university needed to s\_\_\_\_\_ (v.) a balance between the conflicting interests of athletics and academics, guaranteeing players a good education at the expense of winning teams.
2. She went out to the garden for a b\_\_\_\_\_ (n.) of fresh air.
3. He has p\_\_\_\_\_ (adj.) hearing as a result of ear infection (感染).
4. The artist's creativity comes from his k\_\_\_\_\_ (adj.) observation on life and nature.
5. Her d\_\_\_\_\_ (adj.) devotion and love for her country and for her people - her nobility, strength, and courage - are all qualities I aspire to (渴求) as a human being.

## Appendix 10: The worksheet for the think-aloud session

**Directions:** Please answer ONE question from each section. You may use the designated web-based concordancers or any resources available online to help you find the answers. As you do this task, THINK ALOUD your thoughts as naturally and fully as you can. Do not worry about whether your answer is correct, and focus on the process whereby you find the answer.

### I. Choose one INCORRECT option

- \_\_\_\_\_ 1. It may not be easy to \_\_\_\_\_ the concept of globalization.  
(A) grasp (B) seize (C) understand
- \_\_\_\_\_ 2. India's \_\_\_\_\_ economy has brought huge numbers of foreigners and newly affluent (富有的) Indians to the country's big cities.  
(A) booming (B) extending (C) growing
- \_\_\_\_\_ 3. Boating is much more fun than driving and this enjoyment is \_\_\_\_\_ available to vacationers, as live-aboard canal boats are now available for rental.  
(A) easily (B) readily (C) smoothly
- \_\_\_\_\_ 4. The government is urged to \_\_\_\_\_ legislation to protect teenagers against internet pornography.  
(A) introduce (B) pass (C) set
- \_\_\_\_\_ 5. The scientists need to \_\_\_\_\_ data on animal behavior and take blood samples before they can analyze the information to investigate topics such as immunology (免疫系統).  
(A) collect (B) gather (C) pile

### II. Circle one INCORRECT option

- |                    |                   |                   |                 |
|--------------------|-------------------|-------------------|-----------------|
| 1. highly educated | highly exhausted  | highly profitable | highly unusual  |
| 2. vast amounts    | vast figures      | vast majority     | vast numbers    |
| 3. close attention | close exploration | close friend      | close proximity |

### III. Fill in the blank

1. city = u\_\_\_\_\_ area; country = r\_\_\_\_\_ area
2. to use a/the method = to e\_\_\_\_\_ a/the method
3. 簽約 to s\_\_\_\_\_ a/the contract  
毀約 to b\_\_\_\_\_ a/the contract  
終止合約 to t\_\_\_\_\_ a/the contract
4. Death penalty is a c\_\_\_\_\_ issue. Some argue that it violates human rights, while others view it as a necessary evil to prevent crimes.
5. Stress has been found to be a con\_\_\_\_\_ factor in human diseases. It plays a role in triggering or worsening depression and cardiovascular disease (心血管疾病).

IV. Identify and correct the miscollocation in the sentence

For example: He was told to eat the medicine three times a day.  
eat the medicine => take the medicine

1. The meeting aims to attend the issue of violence and crime in America.
2. She was about to step on a major expedition.
3. Labor supply did not increase to achieve demand.
4. The pressure grows up for her as the mid-term exam approaches.

V. Tick the box where the verb collocates with the noun.

Noun \ Verb	E.g., a speed of 60 km/hr	a great success	one's ambition	one's goal	perfection	power	victory
accomplish							
achieve							
attain	V						
gain							

VI. Translation

1. 賺取穩定的收入並非難事。
2. 達不到課程要求的學生將無法畢業。
3. 粗略估計顯示約有八成的受訪者表示滿意這間餐廳的服務。



Appendix 11: The pilot questionnaire

一、基本資料

性別：

男☐

女☐

國籍：

台灣☐

其他(請註明)

母語：

中文☐

其他(請註明)

學習英語的時間：約年

二、英語學習資源

請依據您的實際情形勾選適當的選項

1

你有網路嗎?

是☐

否☐

2

你有用電腦輔助英語學習的經驗嗎?

是☐

否☐

3

你有用過網路字詞檢索工具嗎?

是☐

否☐

4

你有用過搭配詞辭典(collocation dictionary)嗎?

是☐

否☐

5

你常在哪裡使用電腦?(可複選)

家裡☐

學校電腦中心☐

圖書館☐

其他☐

6

你使用網路資源輔助英語學習的頻率為何?

非常頻繁☐

經常☐

有時候☐

很少☐

從未☐

7

你用來輔助學習單字的工具為:(可複選)

一般紙本字典☐

電子字典☐

線上字典☐

其他☐

三、字彙學習經驗

請依據您的學習經驗圈選適當的選項

		6 非常 同意	5 同意	4 傾向 同意	3 傾向 不同意	2 不同意	1 非常 不同意
8	當我遇到生字時，我會從前後文鄰近的字推測字義。	6	5	4	3	2	1
9	當我遇到生字時，我會注意跟這個字搭配出現的字，例如： - sharp 搭配修飾 contrast - painfully 搭配修飾 shy	6	5	4	3	2	1
10	當我遇到生字時，我會從文章上下文的意思推測其字義。	6	5	4	3	2	1
11	學習新字彙時，我會同時學習經常與這個字連用的字。例如 competition： - ‘to face competition’ ‘a fierce/intense competition’	6	5	4	3	2	1
12	學習新字彙時，除了單字本身的意思，我也會去學這個字在句子裡的用法。例如 consist： - The category <b>consists of</b> 3 elements.	6	5	4	3	2	1
13	學習新字彙時，我會去思考這個字跟整篇文章的文義有何關聯性。	6	5	4	3	2	1
14	我學習單字的方法是先背單字的用法、規則，再套用到句子中。	6	5	4	3	2	1



15	我學習單字的方法是先看一些例句，觀察單字在句中的用法，最後再自己歸納出規則來。	6	5	4	3	2	1
16	我學習單字的方法是了解單字的規則，同時觀察單字在例句中的用法，看例句可以幫助我驗證單字的用法是否正確。	6	5	4	3	2	1
17	以下是組成一個單字的要素，身為一個英語學習者，請依照你認為的重要性從 1 到 6 排序 (1: 最重要；6: 最不重要) 拼法 _____ 發音 _____ 字義 _____ 用法規則 _____ 搭配詞 _____ 隱涵的意義 _____ (如 slender 與 skinny 都有'瘦'的意思，但前者意思是褒獎，而後者則有貶抑之意)						
18	你認為網路字詞檢索工具可以幫助你學習字彙的哪些部分? (可複選) 拼法 <input type="checkbox"/> 發音 <input type="checkbox"/> 字義 <input type="checkbox"/> 用法規則 <input type="checkbox"/> 搭配詞 <input type="checkbox"/> 隱涵的意義 <input type="checkbox"/>						

四、字詞搭配知識

請依據您的想法圈選適當的選項

		6 非常 同意	5 同 意	4 傾 向 同 意	3 傾 向 不 同 意	2 不 同 意	1 非 常 不 同 意
19	參與本研究之前，我已經知道'搭配詞'。	6	5	4	3	2	1
20	學習搭配詞可以幫助我選詞用字更加準確。	6	5	4	3	2	1
21	學習搭配詞可以提升我口說或書寫的速度。	6	5	4	3	2	1
22	學習搭配詞可以幫助我根據不同場合選用適當的字詞。	6	5	4	3	2	1
23	網路字詞檢索工具有助於學習搭配詞。	6	5	4	3	2	1
24	本研究中使用的學習單有助於學習搭配詞。	6	5	4	3	2	1
25	這次學習的經驗讓我對搭配詞有更清楚的概念。	6	5	4	3	2	1
26	往後學習英語時，我會特別留意字詞的搭配。	6	5	4	3	2	1
27	我會繼續使用網路字詞檢索工具來輔助我學習搭配詞。	6	5	4	3	2	1

五、數據驅動學習(data-driven learning)與網路字詞檢索工具

請依據你的想法圈選適當的選項

		6 非常 同意	5 同 意	4 傾 向 同 意	3 傾 向 不 同 意	2 不 同 意	1 非 常 不 同 意
28	我認為語料庫中的語言比教科書裡的真實性更高	6	5	4	3	2	1
29	我認為語料庫中的語言比教科書裡的可信度高	6	5	4	3	2	1
30	使用字詞檢索工具搜尋出來的結果數量適中	6	5	4	3	2	1



31	使用字詞檢索工具搜尋出來的結果難易度適中	6	5	4	3	2	1
32	使用字詞檢索工具搜尋出來的結果，大部分我都可以了解	6	5	4	3	2	1
33	雖然大部分搜尋結果都是不完整的句子，但足以供我歸納出關鍵字的搭配詞	6	5	4	3	2	1
34	大量的搜尋結果讓我可以認識到關鍵字的其他搭配詞	6	5	4	3	2	1
35	大量的搜尋結果讓我可以分辨同義字之間的細微差異	6	5	4	3	2	1
36	字詞檢索工具可以幫助我驗證單字的用法是否正確	6	5	4	3	2	1
37	瀏覽大量的搜尋結果可以幫助我對已經學會的字詞印象更深刻	6	5	4	3	2	1
38	用字詞檢索工具搜尋，很難歸納出關鍵字的搭配詞	6	5	4	3	2	1
39	用搜尋結果歸納出搭配詞很耗時	6	5	4	3	2	1
40	搜尋結果數量太多，很難掌握	6	5	4	3	2	1
41	搜尋結果數量太少，難以歸納出搭配詞	6	5	4	3	2	1
42	我覺得由學習者自行用搜尋結果來歸納出單字的用法，有助於強化學習者的自主性	6	5	4	3	2	1
43	我覺得自己歸納出單字的用法，可以加深對單字的印象	6	5	4	3	2	1
44	網路字詞檢索工具的介面清楚	6	5	4	3	2	1
45	網路字詞檢索工具操作容易，使用起來沒有困難	6	5	4	3	2	1
就學習搭配詞而言，你認為網路字詞檢索工具還需要具備哪些功能?(可複選)							
46	清楚、方便操作的介面 <input type="checkbox"/>	搜尋搭配詞的功能 <input type="checkbox"/>					
	語料更多的語料庫 <input type="checkbox"/>	線上字典 <input type="checkbox"/>					
	標明每個字的詞性 <input type="checkbox"/>	單字的出現頻率 <input type="checkbox"/>					
	搜尋結果用完整句子呈現 <input type="checkbox"/>	用影音檔呈現口語語料 <input type="checkbox"/>					



Appendix 12: The questionnaire for the main study (Chinese version)

一、個人基本資料

- 1 性別：

男☐

女☐
- 2 國籍：

台灣☐

其他(請註明)
- 3 母語：

中文☐

其他(請註明)
- 4 學習英語的時間：約 年

二、英語學習資源

請依據您的實際情形勾選適當的選項

5

你有網路嗎?

是☐

否☐

6

你有使用電腦輔助英語學習的經驗嗎?

是☐

否☐

7

參與本研究前，你用過線上字詞檢索工具嗎?

是☐

否☐

8

你用過搭配詞辭典(collocation dictionary)嗎?

是☐

否☐

9

你常在哪裡使用電腦?(可複選)  
家裡☐ 學校電腦中心☐ 圖書館☐ 其他☐

10

你使用網路資源輔助英語學習的頻率為何?  
非常頻繁☐ 經常☐ 有時候☐ 很少☐ 從未☐

11

你用來輔助學習英文單字的工具為:(可複選)  
一般紙本字典☐ 電子字典☐  
線上網路字典☐ 其他☐

三、字彙學習經驗

請依據您的學習經驗圈選適當的選項

		6 非常 同意	5 同意	4 傾向 同意	3 傾向 不同意	2 不同意	1 非常 不同意
12	我遇到生字時，我會從前後鄰近的字推測字義。	6	5	4	3	2	1
13	我遇到生字時，我會從文章上下文的意思推測字義。	6	5	4	3	2	1
14	學習新字彙時，我會注意經常與這個字連用的字。例如 competition: ‘to face competition’; ‘a fierce/intense competition’	6	5	4	3	2	1
15	學習新字彙時，除了單字本身的意思，我也會去學這個字在句子裡的用法。例如 consist: The category <b>consists of</b> 3 elements.	6	5	4	3	2	1
16	學習新字彙時，我會去思考這個單字跟整篇文章的文義有何關聯。	6	5	4	3	2	1
17	我學習單字的方法是先背單字的用法、規則，再套用到句子中。	6	5	4	3	2	1
18	我學習單字的方法是先看一些例句，觀察單字在句中的用法，最後再自己歸納出規則。	6	5	4	3	2	1
19	我學習單字的方法是了解單字的規則，同時觀察單字在例句中的用法，看例句可以幫助我驗證單字的用法是否正確。	6	5	4	3	2	1
	以下是組成一個單字的要素，身為一個英語學習者，請依照你認為的重要性從 1 到 6 排序 (1: 最重要; 6: 最不重要)						
20	拼法 <div></div> 發音 <div></div> 字義 <div></div> 用法規則 <div></div> 搭配詞 <div></div> 隱涵的意義 <div></div> (如 slender 與 skinny 都有‘瘦’的意思，但前者意思是褒獎，而後者則有貶抑之意)						



21	你認為字詞檢索工具可以幫助你學習字彙的哪些部分?(可複選) 拼法 <input type="checkbox"/> 發音 <input type="checkbox"/> 字義 <input type="checkbox"/> 用法規則 <input type="checkbox"/> 搭配詞 <input type="checkbox"/> 隱涵的意義 <input type="checkbox"/>
----	--

四、字詞搭配知識  
請依據您的想法圈選適當的選項

		6 非常 同意	5 同意	4 傾向 同意	3 傾向 不同意	2 不同意	1 非常 不同意
22	參與本研究之前，我已經知道‘搭配詞’。	6	5	4	3	2	1
23	學習搭配詞可以幫助我選詞用字更加精準。	6	5	4	3	2	1
24	學習搭配詞可以提升我口說或書寫的速度。	6	5	4	3	2	1
25	學習搭配詞可以幫助我根據不同場合選用適當的字詞。	6	5	4	3	2	1
26	字詞檢索工具有助於學習搭配詞。	6	5	4	3	2	1
27	本研究使用的學習單有助於學習搭配詞。	6	5	4	3	2	1
28	這次的學習經驗讓我對搭配詞有更清楚的概念。	6	5	4	3	2	1
29	往後學習英語時，我會特別留意字詞的搭配。	6	5	4	3	2	1
30	我會繼續使用字詞檢索工具來輔助我學習搭配詞。	6	5	4	3	2	1

五、線上字詞檢索工具使用心得  
請依據你的想法圈選適當的選項

		6 非常 同意	5 同意	4 傾向 同意	3 傾向 不同意	2 不同意	1 非常 不同意
31	我認為語料庫中的語言比教科書裡的真實性更高	6	5	4	3	2	1
32	我認為語料庫中的語言比教科書裡的可信度更高	6	5	4	3	2	1
33	對我來說，字詞檢索工具搜尋出來的字句數量剛好	6	5	4	3	2	1
34	對我來說，字詞檢索工具搜尋出來的字句難易適中	6	5	4	3	2	1
35	用字詞檢索工具搜尋出來的字句，大部分我都可以了解	6	5	4	3	2	1
36	雖然大部分搜尋結果都是不完整的句子，但足以供我歸納出關鍵字的搭配詞	6	5	4	3	2	1
37	觀察搜尋結果可以幫助我認識關鍵字的其他搭配詞	6	5	4	3	2	1
38	觀察搜尋結果可以幫助我分辨同義字之間的細微差異	6	5	4	3	2	1
39	字詞檢索工具可以幫助我驗證單字的用法是否正確	6	5	4	3	2	1
40	瀏覽大量的搜尋結果可以幫助我對已經學會的字詞印象更深刻	6	5	4	3	2	1
41	用字詞檢索工具搜尋，很難歸納出搭配詞	6	5	4	3	2	1
42	用搜尋結果來歸納出搭配詞很耗時	6	5	4	3	2	1
43	搜尋結果往往數量太多，很難掌握	6	5	4	3	2	1



44	搜尋結果往往數量太少，難以歸納出搭配詞	6	5	4	3	2	1
45	我覺得自己用搜尋結果來歸納出單字的用法，有助於自主學習	6	5	4	3	2	1
46	我覺得自己歸納出單字的用法，可以加深對單字的印象	6	5	4	3	2	1
47	字詞檢索工具的介面清楚	6	5	4	3	2	1
48	字詞檢索工具操作容易，使用起來沒有困難	6	5	4	3	2	1
就學習搭配詞而言，你認為字詞檢索工具需要具備哪些功能?(可複選)							
49	清楚、方便操作的介面 <input type="checkbox"/>	中英雙語的搜尋功能 <input type="checkbox"/>					
	直接搜尋搭配詞的功能 <input type="checkbox"/>	限定詞性的搜尋功能(如 VERB+medicine) <input type="checkbox"/>					
	語料更多的語料庫 <input type="checkbox"/>	搜尋結果用完整句子呈現 <input type="checkbox"/>					
	線上字典 <input type="checkbox"/>	搜尋字的前後字按字母順序排列 <input type="checkbox"/>					
你最喜歡哪個(些)字詞檢索工具，為什麼?(可複選)							
50	Lexical Tutor <input type="checkbox"/>	Collins WordbanksOnline <input type="checkbox"/>					
	NTNU Web Concordancer <input type="checkbox"/>	IWiLL <input type="checkbox"/>					
	TotalRecall <input type="checkbox"/>	VLC Web Concordancer <input type="checkbox"/>					
	請說明原因：						

除了以上問題外，如果您對於本研究或是這種用字詞檢索工具學習搭配詞的學習方式還有任何想法、疑問或心得，請寫在此空格內：



Appendix 13: The questionnaire for the main study (English version)

I. Demographic information

1

Gender

Male ☐

Female ☐

2

Nationality

Taiwan ☐

Other

3

Native language

Chinese ☐

Other

4

Length of studying English

years

II. English learning resources

Please choose the option(s) most applicable to your situation.

5

Do you have access to the Internet?

Yes ☐

No ☐

6

Do you have any experience of computer-assisted language learning?

Yes ☐

No ☐

7

Have you used any web-based concordancer prior to this study?

Yes ☐

No ☐

8

Have you consulted any collocation dictionary?

Yes ☐

No ☐

9

Where do you usually use a computer? (Choose as many as applicable)  
Home ☐ Computer center ☐ Library ☐ Other

10

How often do you use online resources for English learning?  
Always ☐ Often ☐ Sometimes ☐ Seldom ☐ Never ☐

11

What is/are the tool(s) that you use to learn vocabulary? (Choose as many as applicable):  
Conventional dictionaries ☐ Electronic dictionaries ☐  
Online dictionaries ☐ Other  (Please specify)

III. Vocabulary learning experience

Please choose the option most applicable to your situation.

		Strongly agree	Agree	Toward agree	Toward disagree	Disagree	Strongly disagree
12	When I encounter a new word, I infer its meaning from the immediately adjacent words.	6	5	4	3	2	1
13	When I encounter a new word, I infer its meaning from the textual context.	6	5	4	3	2	1
14	When learning a new word, I pay attention to the words that frequently co-occur with the new word. For example, <i>competition</i> : ‘to face competition’; ‘a fierce/intense competition’	6	5	4	3	2	1
15	When learning a new word, I like to know how it is used in a sentence. For example, <i>consist</i> : The category <b>consists of</b> 3 elements.	6	5	4	3	2	1
16	When learning a new word, I like to know how it is used in relation to the text.	6	5	4	3	2	1
17	I learn how to use a word by memorizing its grammatical functions and applying them.	6	5	4	3	2	1
18	I learn how to use a word through reading sample sentences, observing their structures, and then generalizing the grammatical functions from the sentences.	6	5	4	3	2	1
19	I learn how to use a word by knowing its grammatical functions as well as observing sample sentences whereby I can verify the grammatical functions.	6	5	4	3	2	1



20

The following are the components of word knowledge. As an English learner, please arrange the components in the order of your perceived importance (1: most important, 6: least important)

Spelling \_\_\_\_\_ Pronunciation \_\_\_\_\_ Meaning \_\_\_\_\_ Grammatical function \_\_\_\_\_

Collocation \_\_\_\_\_ Connotation \_\_\_\_\_ ('slender' has positive connotation while 'skinny' has negative connotation)

21

Which component(s) of word knowledge do you think concordancers facilitate? (Choose as many as applicable)

Spelling ☐ Pronunciation ☐ Meaning ☐ Grammatical function ☐ Collocation ☐

Connotation ☐

IV. Collocational awareness

Please choose the option most applicable to your situation.

		Strongly agree	Agree	Toward agree	Toward disagree	Disagree	Strongly disagree
22	I have already known the concept of <i>collocation</i> prior to this study.	6	5	4	3	2	1
23	Learning collocations enhances the accuracy of word choice.	6	5	4	3	2	1
24	Learning collocations enhances the fluency of written or spoken production.	6	5	4	3	2	1
25	Learning collocations enhances the appropriateness of word choice according to register/genre.	6	5	4	3	2	1
26	The concordancers are helpful in learning collocations.	6	5	4	3	2	1
27	The collocation learning worksheets are helpful in learning collocations.	6	5	4	3	2	1
28	This learning experience raises my awareness of collocation.	6	5	4	3	2	1
29	I will pay attention to collocations in future English learning.	6	5	4	3	2	1
30	I will keep using web-based concordancers to assist my collocation learning.	6	5	4	3	2	1

V. Experience of learning collocations with web-based concordancers

Please choose the option(s) most applicable to your situation.

		Strongly agree	Agree	Toward agree	Toward disagree	Disagree	Strongly disagree
31	I think the language in corpora is more authentic than that in coursebooks.	6	5	4	3	2	1
32	I think the language in corpora is more credible than that in coursebooks.	6	5	4	3	2	1
33	I find the amount of concordancing output adequate.	6	5	4	3	2	1
34	I find the difficulty level of concordancing output adequate.	6	5	4	3	2	1
35	I understand most of the concordancing output.	6	5	4	3	2	1
36	Although most concordancing output is displayed in incomplete sentences, it is sufficient to generalize frequent collocations of the target word.	6	5	4	3	2	1
37	Observing concordancing output allows me to discover a range of collocates of the target word.	6	5	4	3	2	1
38	Observing concordancing output helps me to notice the nuances between near-synonyms.	6	5	4	3	2	1



39	Concordancers help me verify my linguistic hypotheses.	6	5	4	3	2	1
40	Corpus consultation consolidates the knowledge of the words studied previously.	6	5	4	3	2	1
41	It is difficult to generalize collocations from corpus data.	6	5	4	3	2	1
42	It is time-consuming to generalize collocations from corpus data.	6	5	4	3	2	1
43	The amount of search results is too large to manage.	6	5	4	3	2	1
44	The amount of search results is too small to generalize collocations.	6	5	4	3	2	1
45	Generalizing collocations from corpus data enhances my sense of autonomous learning.	6	5	4	3	2	1
46	Corpus consultation enhances the knowledge of the words studied previously.	6	5	4	3	2	1
47	The interfaces of the web-based concordancers are user-friendly.	6	5	4	3	2	1
48	The web-based concordancers are easy to use.	6	5	4	3	2	1
What other feature(s) do you think a web-based concordancer should provide to better assist collocation learning? (Choose as many as applicable)							
49	<div> <div> user-friendly interface <input type="checkbox"/></div> <div>large corpus size <input type="checkbox"/></div> <div>bilingual search <input type="checkbox"/></div> <div>results displayed in complete sentences <input type="checkbox"/></div> </div> <div> <div>collocation search <input type="checkbox"/></div> <div>online dictionary <input type="checkbox"/></div> <div>part-of-speech search <input type="checkbox"/></div> <div>sorting function <input type="checkbox"/></div> </div>						
Which web-based concordancer(s) do you prefer, and why? (Choose as many as applicable)							
50	<div> <div>Lexical Tutor <input type="checkbox"/></div> <div>NTNU Web Concordancer <input type="checkbox"/></div> <div>TOTALrecall <input type="checkbox"/></div> </div> <div> <div>Collins WordbanksOnline <input type="checkbox"/></div> <div>IWiLL <input type="checkbox"/></div> <div>VLC Web Concordancer <input type="checkbox"/></div> </div>						
Please provide a rationale for your choice(s):							

Please leave your comments in this box if you have further opinions, queries or thoughts about this study or the pedagogical approach to learning collocations assisted by concordancers.



Appendix 14: Descriptive statistics for questionnaire responses

I. Demographic information

1	Gender	Male = 13	Female = 85
2	Nationality	Taiwan = 98	Other = 0
3	Native language	Chinese = 98	Other = 0
4	Length of studying English	M=8.12 (yrs)	

II. English learning resources

5	Do you have access to the Internet?	Yes = 98	No = 0
6	Do you have any experience of computer-assisted language learning?	Yes = 84	No = 14
7	Have you used any web-based concordancer prior to this study?	Yes = 39	No = 59
8	Have you consulted any collocation dictionary?	Yes = 14	No = 84
9	Where do you usually use a computer? (Choose as many as applicable) Home = 98   Computer center = 32   Library = 5   Other = 1		
10	How often do you use online resources for English learning? Always = 5   Often =26   Sometimes =55   Seldom =10   Never =1		
11	What is/are the tool(s) that you use to learn vocabulary? (Choose as many as applicable): Conventional dictionaries =51   Electronic dictionaries =79 Online dictionaries = 90   Other =2		

III. Vocabulary learning experience

		Strongly agree	Agree	Toward agree	Toward disagree	Disagree	Strongly disagree
12	When I encounter a new word, I infer its meaning from the immediately adjacent words.	17	48	29	3	1	0
13	When I encounter a new word, I infer its meaning from the textual context.	24	50	21	3	0	0
14	When learning a new word, I pay attention to the words that frequently co-occur with the new word. For example, <i>competition</i> : ‘to face competition’; ‘a fierce/intense competition’	5	20	56	13	1	0
15	When learning a new word, I like to know how it is used in a sentence. For example, <i>consist</i> : The category <b>consists of</b> 3 elements.	14	37	38	8	1	0
16	When learning a new word, I like to know how it is used in relation to the text.	7	30	48	10	1	0
17	I learn how to use a word by memorizing its grammatical functions and applying them.	6	41	38	11	2	0
18	I learn how to use a word through reading sample sentences, observing their structures, and then generalizing the grammatical functions from the sentences.	9	28	43	16	2	0
19	I learn how to use a word by knowing its grammatical functions as well as observing sample sentences whereby I can verify the grammatical functions.	17	33	41	5	2	0



- 20 The following are the components of word knowledge. As an English learner, please arrange the components in the order of your perceived importance (1: most important, 6: least important).  
Spelling M=3.94 Pronunciation M=3.37 Meaning M=2.20 Grammatical function M=3.31  
Collocation M=4.36 Connotation M=4.39

- 21 Which component(s) of word knowledge do you think concordancers facilitate? (Choose as many as applicable)  
Spelling =22 Pronunciation =18 Meaning =52 Grammatical function =83  
Collocation =78 Connotation =28

#### IV. Collocational awareness

		Strongly agree	Agree	Toward agree	Toward disagree	Disagree	Strongly disagree
22	I have already known the concept of <i>collocation</i> prior to this study.	11	29	25	23	7	3
23	Learning collocations enhances the accuracy of word choice.	29	37	28	3	1	0
24	Learning collocations enhances the fluency of written and/or spoken production.	10	28	46	13	1	0
25	Learning collocations enhances the appropriateness of word choice according to register/genre.	22	35	37	3	1	0
26	The concordancers are helpful in learning collocations.	30	38	27	2	1	0
27	The collocation learning worksheets are helpful in learning collocations.	24	42	26	5	0	0
28	This learning experience raises my awareness of collocation.	21	42	30	5	0	0
29	I will pay attention to collocations in future English learning.	24	42	28	3	0	0
30	I will keep using web-based concordancers to assist my collocation learning.	29	26	36	5	2	0

#### V. Experience of learning collocations with web-based concordancers

		Strongly agree	Agree	Toward agree	Toward disagree	Disagree	Strongly disagree
31	I think the language in corpora is more authentic than that in coursebooks.	6	25	55	11	1	0
32	I think the language in corpora is more credible than that in coursebooks.	7	16	60	15	0	0
33	I find the amount of concordancing output adequate.	2	16	40	38	2	0
34	I find the difficulty level of concordancing output adequate.	3	17	61	15	1	0
35	I understand most of the concordancing output.	4	20	61	10	3	0
36	Although most concordancing output is displayed in incomplete sentences, it is sufficient to generalize frequent collocations of the target word.	10	30	46	10	2	0
37	Observing concordancing output allows me to discover a range of collocates of the target word.	12	42	41	3	0	0
38	Observing concordancing output helps me to notice the nuances between near-synonyms.	3	33	44	17	1	0
39	Concordancers help me verify my linguistic hypotheses.	14	38	37	8	0	0
40	Corpus consultation consolidates the knowledge of the words studied previously.	8	32	45	10	1	2
41	It is difficult to generalize collocations from corpus data.	3	9	32	51	3	0



42	It is time-consuming to generalize collocations from corpus data.	9	11	35	36	6	0
43	The amount of search results is too large to manage.	8	16	43	27	3	0
44	The amount of search results is too small to generalize collocations.	6	14	37	37	4	0
45	Generalizing collocations from corpus data enhances my sense of autonomous learning.	10	30	48	9	0	0
46	Corpus consultation enhances the knowledge of the words studied previously.	13	35	44	5	0	0
47	The interfaces of the web-based concordancers are user-friendly.	3	18	42	28	4	1
48	The web-based concordancers are easy to use.	5	22	45	22	1	1
What other feature(s) do you think a web-based concordancer should provide to better assist collocation learning? (Choose as many as applicable)							
49	<div> <div> user-friendly interface =87  large corpus size =74  bilingual search =64  results displayed in complete sentences =54 </div> <div> collocation search =67  online dictionary =68  part-of-speech search =74  sorting function =52 </div> </div>						
Which web-based concordancer(s) do you prefer, and why? (Choose as many as applicable)							
50	<div> <div> Lexical Tutor =5  NTNU Web Concordancer =13  TOTALrecall =12 </div> <div> Collins WordbanksOnline =65  IWiLL =39  VLC Web Concordancer =13 </div> </div>						



## Appendix 15: Restricted collocations in writing assignments (experimental group)

	Pre-treatment	Post-treatment
verb + noun	apply ointment come into one's eyes find time have a good time hold a party make bed spent time surf the Internet take a bath/shower take up space watch DVD/TV/TV program	break a habit cause a problem/ procrastination comes into mind dive in a task encounter a challenge/ difficulty/ problem/obstacle face difficulty/failure/fear/problem /trouble/task/work follow a direction/idea/method /plan/schedule/step give sb a hand give sth a try go to school/university hang out with friends have fun join a band learn a lesson make a beginning/decision/plan make friends meet difficulty/task plays a role play badminton/basketball/tennis /volleyball solve a problem spend/take time take a break/rest/trip take a breath take a bus take a photo/picture tell the difference watch a comedy/game/ local drama /movie/ soap opera/TV/program
noun + verb	earthquake hits (old) saying goes smell floats	dream comes true proverb/saying goes
adjective + noun	big meal poor handwriting warm yellow	big fan/lunch/problem/trouble big difference/reason/task bright/dark side free time full consideration heavy task/work psychological barrier right path tacit understanding
noun of noun	Nil	Nil
adverb+ adjective	Nil	Nil
verb+ adverb	fall asleep believe ... deeply	love ... deeply

## Appendix 16: Restricted collocations in writing assignments (control group)

	pre	post
verb + noun	absorb knowledge abuse freedom achieve a goal answer needs attach importance build a foundation/relationship cultivate a habit cut/join/pass/skip a class encounter ... situation enter college/school/society/university exhaust one's energy face a problem fail/pass/take a course fight for one's rights find a job follow direction/progress/rule/schedule /step go to college/school/university grasp a chance have fun join a club lead a life make arrangement/choice/decision /difference/progress make money make friends obey/observe a rule pass/take an exam pay attention place restrictions play a part/role reach a goal solve a problem spend/take time take a break/trip take the consequence/control take a lesson take notes watch TV	absorb information/knowledge attach importance bear pressure belong to a club bring up a question carry out a policy/project cut down time do one's best enter college expand interests extend relationship fail/pass/take a course face challenge/question/pressure/task find a job follow the principle go to college/school have fun hold an activity join an association/club/team join in an activity make a mistake make friends make progress pass a subject pay attention play a role play guitar/piano play music reach a consensus reach a level resolve/solve a problem relieve burden/pressure run affairs/clubs save/spend/take time skip a class suffer hardships surf the Internet take a class take a break/rest/trip take the consequence upset a plan watch TV
noun + verb	time goes by	proverb says
adjective + noun	big burden deep impression free time gloomy life strong impact	slow learner heavy burden/homework/pressure free time rote learning full schedule deep impression big challenge/problem/trouble hot issue poor education/grade virtual life depressed economy



noun of noun	Nil	Nil
adverb+ adjective	Nil	Nil
verb+ adverb	fall asleep	fall asleep

# **Appendix 17: Miscollocations in writing assignments (experimental group)**

	Pre	Post
Literal translation	affirmatively say concentrated reading lightly full long-eager dream outside bothers/scenery stupefied face sunshine sprays/sprinkles to arrange a bed/bookcase to sit a position to walk a road to write down data uncomfortable smell	retarded way skilful subject small fulfilment tight nerves to adjust one's mind to chat a topic to handle a reason to influence friendship to obey one's idea
Deviant translation	confused bed higgledy-piggledy desk indiscriminate room/waste yard noble lamp omnibearing room (a) unrestrained fly garbage diffuses sleep sedately table disseminates something tender light to knead one's head to lay bare silk stockings to transpire a smell	Nil
Using synonyms	blond door moist towel overdue food to view the sea	bright skin faulty method nice price to examine homework to own a hobby to pay time
Analogy	to take a hot spring	Nil
Paraphrase	a crowd of cookies a piece of dust automatic kitchen/room check my dressing clean the outcome collapse chaotically complex room constant outflow cute curtain decayed snacks dingy bed disarranged pillow/quilt disgusting odor disordered bed/blanket disorderly bag/store disused bathroom dreadful mudflow failed report card fall irregularly fan dangles lovely style	a piece of task achieve work amazing destiny good wisdom calm down one's mind conquer a task/work cut in a task deal with sth blindly declare an opinion deficient homework delightful mood dilatory people eradicate procrastination face an experience fast human fight with a task finish a goal finish instantly funny personality happy emotion jump into a problem



	messily laid messy layout mysterious desk neat shoes/socks organized room overflowing number secret zone shelf hangs sloping curtain rod spread a plug suitable wearing whole sight yucky liquid	left personality behind major in a department meet a project overcome a task put off one's duty release tiredness resist distraction/fear/task solve behaviour/habit/procrastination /task terminate a habit throw off procrastination wear casually
Using de-lexicalized words	make a damage/travel take on clothes	do a plan get achievement/difference/guts /progress have procrastination set a reward set up confidence
Synformy	clear bedclothes relative photo to clear bed	adopt in the environment
Grammatical error	Nil	procrastinate the thing/work
Unintelligible combination	nightclothes have no plait unlashing clothes	Nil

## Appendix 18: Miscollocations in writing assignments (control group)

	Pre	Post
verb + noun	<p> abandon a class  accept education  acquire learning/patience/result  adapt to boss/colleague  affect accomplishment  align with one's interest  arrange the class/future  assign the curriculum  attend the office  attract concentration  board the erudition/relationship  catch a change  challenge a class  conflict with a course  conform to one's interest  confront tomorrow  cultivate competence  defeat sleepiness  delve into a skill  describe an excuse  develop brain potency  dig into a thing  discern the experience  dissipate money/time  do a decision/responsibility  extend the mind/viewpoint  feel sleep deprivation  fight the professor  figure out a subject  gain learning  get a course/definition  get rid of stress  go to truant  handle behaviour  have absence/ concentration  / crime (in heart)  inspire the potential  join an action  know general sense/knowledge  learn erudition/course/scholarship  listen the curriculum  lose knowledge/interpersonal  relationship  make a boyfriend  make a direction  manipulate time  master progress  open the web cam  overcome a student  own a choice/ freedom/opportunity  pick up a class/course  practice the gold/social contact  prepare the competence  raise an ability  realize the direction/mind </p>	<p> absorb a course  accomplish/achieve homework  accumulate strength  add burden  adjust a viewpoint  advance the ability/friendship  afford homework  attend an association  attend the exam  catch the content/experience/skill  close eyebrow  conduct an assignment  confront the pressure  convey to a question  cost time  cultivate a hobby  distribute time  enrich the report  expand relationship  forfeit the opportunity  forget the curriculum  get melancholia/understanding  grab information/skill/university  grasp learning  improve a sense of achievement  increase laziness  inquire knowledge  level up ability  lighten a financial problem  listen to the theory  look up the Internet  lose the purpose/zeal  make learning/thoughts  motivate the desire  obtain a concept  own the accomplishment/time  pass a concept  pay a report  poison enthusiasm  prevent the gap/laziness  reach the quality  realize the concept/course/jargon  / knowledge  receive learning  remember knowledge  rise learning  satisfy the quality/quantity  score the work  seek a goal  solve frictions/tasks  stand in the society  study knowledge  support livelihood  take extra-curricular activity  take part in an association </p>



verb+noun (contd.)	recognize friends refuse the attraction reply an answer say affairs shoulder the result side with the idea solve a condition speak one's opinion stand the consequence/job stipulate a student take place a field trip teach erudition/progress tend a concept fit the thought remain a good impression touch knowledge treasure a chance/opportunity	tell homework temporize to professors track a answer train competitiveness/responsibility /a sense of responsibility treat one's aspect uphold in learning wonder more time
noun + verb	spirit stays	chapter pulls up competitiveness shrinks word happens
adjective + noun	active attitude autonomous competence conflicted schedule elastic timetable heavy self-awareness irresistible reason lazy personality quiet occasion ripe student skilful knowledge strong assistance substantial life supportive view surplus time varied erudition voluntary learning	bad comprehension big stress conjoined time complete sleep concrete class depressed reason different comprehension emotional reflection grand activity happy surrounding high competitiveness ineffective consequence irresistible homework large homework naïve emotional reflection narrow learning oncoming society passive education perfunctory chance/student proper pressure/quantity smooth atmosphere substantial learning unceasingly practice unified time unitary learning
noun of noun	Nil	Nil
adverb+ adjective	easily distrait	Nil
verb+ adverb	aggravate easily attend class earnestly learn earnestly read cursorily stand dumbly study conscientiously/voluntarily	learn spontaneously/voluntarily study spontaneously/voluntarily

Appendix 19: Inferential statistics for Questionnaire Items 20, 21 and 50

1. Item 20 (results of Friedman’s ANOVA and Wilcoxon signed-rank test)

Ranks		Test Statistics <sup>a</sup>	
	Mean Rank		
Q20拼法	3.86	N	95
Q20發音	3.27		
Q20字義	2.09		
Q20用法	3.21		
Q20搭配詞	4.29		
Q20隱涵	4.28	Chi-Square	97.310
		df	5
		Asymp. Sig.	.000
a. Friedman Test			

		Test Statistics														
		Q20發音	Q20字義	Q20用法	Q20拼法	Q20搭配詞	Q20隱涵	Q20字義	Q20發音	Q20用法	Q20發音	Q20拼法	Q20搭配詞	Q20隱涵	Q20字義	Q20發音
Z		-2.441 <sup>a</sup>	-6.013 <sup>a</sup>	-2.387 <sup>a</sup>	-1.767 <sup>b</sup>	-1.564 <sup>b</sup>	-4.177 <sup>a</sup>	-.371 <sup>a</sup>	-3.783 <sup>b</sup>	-3.443 <sup>b</sup>	-5.995 <sup>b</sup>	-7.672 <sup>b</sup>	-7.014 <sup>b</sup>	-6.317 <sup>b</sup>	-4.547 <sup>b</sup>	-.490 <sup>b</sup>
Asymp. Sig. (2-		.015	.000	.017	.077	.118	.000	.711	.000	.001	.000	.000	.000	.000	.000	.624
a. Based on positive ranks.																
b. Based on negative ranks.																
c. Wilcoxon Signed Ranks Test																



2. Item 21 (results of Cochran's test and McNemar's test)

Frequencies

	Value	
	0	1
Q21拼法	76	22
Q21發音	80	18
Q21字義	46	52
Q21用法	15	83
Q21搭配詞	20	78
Q21隱涵	70	28

Test Statistics

N	98
Cochran's Q	162.834 <sup>a</sup>
df	5
Asymp. Sig.	.000

a. 1 is treated as a success.

Test Statistics<sup>a</sup>

	Q21拼法 & Q21發音	Q21拼法 & Q21字義	Q21拼法 & Q21用法	Q21拼法 & Q21搭配詞	Q21拼法 & Q21隱涵	Q21發音 & Q21字義	Q21發音 & Q21用法	Q21發音 & Q21搭配詞	Q21發音 & Q21隱涵	Q21字義 & Q21用法	Q21字義 & Q21搭配詞	Q21字義 & Q21隱涵	Q21用法 & Q21搭配詞	Q21用法 & Q21隱涵	Q21搭配詞 & Q21隱涵
N	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98
Chi-Square <sup>a</sup>		20.024		42.014	.694	27.225	53.195	44.628	2.250	19.149	11.161	10.580	51.158	48.020	
Asymp. Sig.		.000		.000	.405	.000	.000	.000	.134	.000	.001	.001	.000	.000	
Exact Sig. (2-tailed)	.523 <sup>b</sup>												.359 <sup>b</sup>		

- a. Continuity Corrected  
b. Binomial distribution used.  
c. McNemar Test

3. Item 50 (results of Cochran's test and McNemar's test)

Frequencies

	Value	
	0	1
Q50Lexical	93	5
Q50NTNU	85	13
Q50Total	86	12
Q50Collins	33	65
Q50Iwill	59	39
Q50VLC	85	13

Test Statistics

N	98
Cochran's Q	131.046 <sup>a</sup>
df	5
Asymp. Sig.	.000

a. 0 is treated as a success.

Test Statistics<sup>c</sup>

	Q50Lexical & Q50NTNU	Q50Lexical & Q50Total	Q50Lexical & Q50Collins	Q50Lexical & Q50Iwill	Q50Lexical & Q50VLC	Q50NTNU & Q50Total	Q50NTNU & Q50Iwill	Q50NTNU & Q50VLC	Q50Total & Q50Collins	Q50Total & Q50Iwill	Q50Total & Q50VLC	Q50Collins & Q50Iwill	Q50Collins & Q50VLC	Q50Iwill & Q50VLC
N	98	98	98	98	98	98	98	98	98	98	98	98	98	98
Chi-Square <sup>a</sup>				32.029					41.600	15.721		9.191	43.350	12.500
Asymp. Sig.				.000					.000	.000		.002	.000	.000
Exact Sig. (2-tailed)	.077 <sup>b</sup>	.118 <sup>b</sup>	.000	.077 <sup>b</sup>				1.000 <sup>b</sup>			1.000 <sup>b</sup>			

a. Continuity Corrected

b. Binomial distribution used.

c. McNemar Test